MODELING SMALL UNMANNED AERIAL SYSTEM MISHAPS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

THESIS

Sean E. Wolf, Captain, USAF

AFIT-OR-MS-ENS-12-29

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-OR-MS-ENS-12-29

MODELING SMALL UNMANNED AERIAL SYSTEM MISHAPS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Sean E. Wolf, BS, MA

Captain, USAF

March 2012

AFIT-OR-MS-ENS-12-29

MODELING SMALL UNMANNED AERIAL SYSTEM MISHAPS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

Sean E. Wolf, BS, MA
Captain, USAF

Approved:

________//SIGNED//__________________________ 19/03/2012___
Dr. Raymond R. Hill, Jr. (Chairman) date

_________//SIGNED//_________________________ 19/03/2012___
Dr. Joseph J. Pignatiello, Jr. (Member) date

AFIT-OR-MS-ENS-12-29

Abstract

A dataset of 854 small unmanned aerial system (SUAS) flight experiments from 2005-2009 is analyzed to determine significant factors that contribute to mishaps. The data from 29 airframes of different designs and technology readiness levels were aggregated. Twenty measured parameters from each flight experiment are investigated, including wind speed, pilot experience, number of prior flights, pilot currency, etc. Outcomes of failures (loss of flight data) and damage (injury to airframe) are classified by logistic regression modeling and artificial neural network analysis.

From the analysis, it can be concluded that SUAS damage is a random event that cannot be predicted with greater accuracy than guessing. Failures can be predicted with greater accuracy (38.5% occurrence, model hit rate 69.6%). Five significant factors were identified by both the neural networks and logistic regression.

SUAS prototypes risk failures at six times the odds of their commercially manufactured counterparts. Likewise, manually controlled SUAS have twice the odds of experiencing a failure as those autonomously controlled. Wind speeds, pilot experience, and pilot currency were not found to be statistically significant to flight outcomes. The implications of these results for decision makers, range safety officers and test engineers are discussed.

Acknowledgments

I would like to express sincere appreciation to my faculty advisor, Dr. Ray Hill, for his encouragement, guidance, and sense of adventure for taking me on before the full scope of this project was well defined. I would also like to thank my reader, Dr. Joseph Pignatiello, for his technical insights and helpful comments. Thanks also to Dr. Ken Bauer for the challenging (but valuable) multivariate analysis course and for instructing me in neural network screening methods. Thanks to 2d Lt Harris Butler for the technical MATLAB assistance that allowed me to put Dr. Bauer's methods into practice. Thanks to Lt Col Anthony Tvaryanas for insights into the (often non-technical) finer points of large-scale mishap investigations.

Big thanks are due my sponsor, Mr. Johnny Evers at AFRL/RWWV, for releasing the flight data for my analysis and encouraging me to pursue this mutually beneficial research. Many thanks are due Mr. Ken Blackburn for his subject-matter expertise and his helpful comments as my research progressed. Special thanks to my father for keeping my two-year-old son entertained while I worked long hours finishing this thesis. Lastly, and most importantly, I am indebted to my wife for her loving support, for her endless devotion, and for providing me with the most brutally honest sanity checks on my work I could ever have hoped for.

Sean E. Wolf

Table of Contents

Page

List of Figures

Page

Page

List of Tables

Page

List of Acronyms

AFB	Air Force Base
AFRL	Air Force Research Laboratory
ANCOVA	Analysis of Covariance
ANN	Artificial Neural Network
AUC	Area Under the Curve
BAO	Battlefield Air Operations
COA	Certificate of Authorization
COTS	Commercial Off The Shelf
CRM	Crew Resource Management
DoD	Department of Defense
FAA	Federal Aviation Administration
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis
HFACS	Human Factors Analysis and Classification System
MTBF	Mean Time Between Failures
NAS	National Air Space
NASA	National Aeronautics and Space Administration
OLS	Ordinary Least Squares
ORM	Operational Risk Management
OSD	Office of the Secretary of Defense

R/C Remote Control

ROC Receiver Operating Characteristic

SAA Sense and Avoid

SNR Signal-to-Noise Ratio

SUAS Small Unmanned Aerial System

TALS Tactical Automated Landing System

UAS Unmanned Aerial System

UAV Unmanned Aerial Vehicle

MODELING SMALL UNMANNED AERIAL SYSTEM MISHAPS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

I. Introduction

Small Unmanned Aerial Systems (SUAS) are proliferating throughout the armed forces, law enforcement and civilian sectors. There are tens of thousands of SUAS in service around the world, comprising hundreds of unique airframes used for dozens of diverse missions. Miniaturization, improvements in autopilot technology and the development of advanced batteries have enabled SUAS to flourish where once only larger Unmanned Aerial Systems (UAS) were feasible.

This explosion in the SUAS population has meant great gains for military units who now can quickly employ a cheap reconnaissance platform without risking a pilot, or an expensive aircraft. However, UAS in general, both large and small, tend to be much less reliable than manned systems. The extent of current UAS analysis has been limited to large systems, and the results of that analysis are not encouraging. Large UAS across all platforms and services have historically seen mishap rates one to two orders of magnitude higher than manned aircraft (OSD 2009).

Reliability is a critical issue for all UAS because “it underlies their affordability (an acquisitions issue), their mission availability (an operations and logistics issue), and their acceptance into civil airspace (a regulatory issue)” (OSD 2003). Given the dearth of

data for SUAS, organizations like the Federal Aviation Administration (FAA) are hesitant to grant Certificates of Authorization (COAs) for SUAS flight in the National Airspace (NAS). Research organizations like the Air Force Research Laboratory (AFRL) must make important acquisition and flight testing decisions about this often unpredictable technology, putting money and flight test safety at risk in the process. Operational units purchase and fly SUAS platforms, putting their mission effectiveness and troop safety in the hands of a technology with little published data. With data on SUAS reliability, informed decisions could be made across the spectrum of SUAS operations, from the regulatory side through development, test and evaluation, to operational deployment of these systems. With an understanding of the unique nature of SUAS and insight into the causes of their mishap rates, millions of dollars could potentially be saved throughout the acquisitions lifecycle of this technology.

This thesis uses a dataset of SUAS flights from AFRL's Munitions Directorate to ascertain the root causes of SUAS mishaps to exploit them for process improvement and lead to future mishap prevention. AFRL flies over two dozen types of SUAS with wingspans from 20 inches to 11 feet and weights from one to 100 pounds. They use a mixture of electric and gasoline propulsion. AFRL's SUAS fleet represents a wide swath of the sizes, payloads and propulsion types found in the general SUAS population. The dataset that AFRL provided for this analysis is composed of five years' worth of SUAS experimental flight testing (from 2005-2009) with over 850 unique flights, 29 unique airframes and 103 different tail numbers. The results of each flight were recorded in flight reports and root causes were identified or hypothesized for all mishaps and aircraft damage. In all, 19 unique parameters were extracted or derived from the flight reports,

including surface wind speed, ambient temperature, pilot's previous number of flights, days since airframe last flown, wingspan of airframe, and time of day flown.

This thesis utilizes multivariate data analysis techniques to attempt to classify flights by mishap potential based on AFRL's historical records and the parameters that can be obtained prior to flight. Logistic regression is employed to develop classification functions and to quantity the impact of key factors on mishaps. Artificial neural network feature screening techniques are utilized to identify the most significant factors for classifying SUAS mishaps so that they can be investigated for process improvement. The root causes of SUAS mishaps are then exploited to create mishap prevention strategies. Existing mishap prevention strategies for large UAS are considered and analyzed for their potential applicability to SUAS in light of the mishap factors identified by this analysis.

II. Literature Review

To date, there have been no published statistics on SUAS reliability, although in the past 10 years, some reports on UAS reliability have been generated for larger platforms. An explanation for this lack of detail in early research was offered by the FAA in 2004: "[military UAS] are much less expensive than manned aircraft and so do not warrant the same level of analysis" (Williams 2004). That may have been true in 2004 but today, when the military services are spending hundreds of millions of dollars acquiring SUAS, the justification for further analysis is clear.

Mishap Reports

The primary mechanism by which to track large UAS reliability is via mishap reports. Mishap reports document incidents in which an aircraft caused unintended damage exceeding a certain dollar amount or injuries to friendly personnel or noncombatants. As Nullmeyer, Herz and Montijo (2009) point out, "It is clear that mishap frequencies, rates and causes are all dynamic in the emerging field of UAS operations, and that mishap reports provide a fertile source of insight into where training and operations need to be improved."

Mishap classification in the Department of Defense is governed by DoD Instruction 6055.07, "Mishap Notification, Investigation, Reporting, and Record Keeping". This document defines responsibilities and procedures for mishap

investigations and provides the classification scheme to be used by the component services. DoDI 6055.07 lists the following mishap classifications:

> <u>Class A mishap.</u> The resulting total cost of damages to Government and other property is $2 million or more, a DoD aircraft is destroyed (excluding UAS Groups 1, 2, or 3), or an injury or occupational illness results in a fatality or permanent total disability.
>
> <u>Class B mishap.</u> The resulting total cost of damages to Government and other property is $500,000 or more, but less than $2 million. An injury or occupational illness results in permanent partial disability, or when three or more personnel are hospitalized for inpatient care (which, for mishap reporting purposes only, does not include just observation or diagnostic care) as a result of a single mishap.
>
> <u>Class C mishap.</u> The resulting total cost of property damages to Government and other property is $50,000 or more, but less than $500,000; or a nonfatal injury or illness that results in 1 or more days away from work, not including the day of the injury.
>
> <u>Class D mishap.</u> The resulting total cost of property damage is $20,000 or more, but less than $50,000; or a recordable injury or illness not otherwise classified as a Class A, B, or C mishap.

Maintenance records and flight logs are not generally accessible for analysis, but mishap statistics are collected and published by the different branches of the military. The mishap reports generated from these events for large UAS have been collected and analyzed by several scholars who sort and group the mishap causes into different classifications.

Mishap Factors

A consensus opinion to emerge from analysis of the data is that large UAS have a much higher mishap rate than manned aircraft (Williams 2004). This has been attributed to numerous factors. Human error was the most often cited cause. In early studies, it was found to comprise anywhere from 21% to 80% of all mishaps (Williams 2004). More recent studies have found that human error is a mishap cause in a range between 56-69% of all mishaps (Tvaryanas and Thompson 2008). The other mishap factors are often lumped under general categories, like “engine” or “structure” for those cases when a cause has been determined at all.

The mishap factors varied in extent by aircraft. Given that the different branches of the military fly differing UAS, the mishap rates varied by service. The difficulty in comparing these human factors mishap rates across systems was summarized well by Williams: “[M]ost of the other human factors-related accidents were unique in the sense that a problem that occurred for one type of aircraft would never be seen for another because the user interfaces for the aircraft are totally different” (Williams 2004).

The majority of research into the causes of these mishaps has focused on the human factors involved. This is because engineering solutions are expected to progress as they have for manned systems and gradually yield lower UAS mishap rates with system maturation (Nullmeyer, Herz and Montijo 2009). Indeed, optimism has been expressed that these engineering and automation improvements would lead to reduced human factors errors as well: “The effect of human error is expected to decrease as the level of autonomy increases and operators gain more experience” (Dalamagkidis, Valavanis and

Piegl 2008). These improvements are expected to occur over time, as they do for all new technologies, therefore, the majority of literature on UAS mishaps has concentrated on human factors, which is viewed as an area that can be immediately exploited for process improvement.

An overlap between the human factors and technical causes of mishaps is that of time, usually measured in number of flight hours. UAS safety performance is expected to improve in most measures given more time to learn the intricacies of these complex systems. Failure rates should be nonlinear and decreasing after “increased experience in the operation of a given UAS type” (Clothier, et al. 2011). Additionally, OSD reports that large UAS have seen improvements in mishap rates over recent years, with their measured “reliability approaching an equivalent level of reliability to their manned military counterparts” (OSD 2009). OSD expects, therefore, that large UAS mishap rates will improve over time, specifically due to “flight experience” and “improved technologies” (OSD 2009). Time is thus expected to correlate with increased human performance and decreased technical risks.

Technical Risks and Reliability

Researchers have hypothesized other technical risks to manned and unmanned aircraft operations that may not be time- or learning-curve-dependent, including atmospheric conditions and maintenance reports. For UAS, NASA’s experience has shown that “the most important operational consideration for flight has become the weather” (Teets, et al. 1998). Specifically within weather considerations, NASA found

wind speed and direction to be the most important meteorological consideration (Teets, et al. 1998). While the above assertions are based on NASA's experience, and support their call for better atmospheric data characterization, no data were provided to quantify the effect of climate on UAS performance. Quantifiable research by the US Air Force has considered the effect of average surface temperature at a pilot's home base as a potential mishap factor for manned aircraft. The results revealed "no significant statistical correlation between extreme surface temperatures at home station and the flight mishap rates" (Miarecki and Constable, 2007). Likewise, Marine Corps monthly maintenance reports were analyzed to determine if their contents could predict future AV-8 Harrier mishaps, but no statistically significant model was found (Van Houten 1994). While these two empirical results pertain to manned aircraft, each address important factors to consider for SUAS, although no comparable studies for UAS of any size have been found.

Some studies of SUAS reliability have considered Fault Tree Analysis (FTA) and Failure Modes and Effects Analysis (FMEA). Each involve engineering practices where the system is defined as subsystems or components and their individual reliabilities are analyzed to determine likelihood of faults and their resulting risk scenarios. Cline (2008) attests that FTA and FMEA serve as useful tools for determining levels of SUAS reliability and Dermentzoudis (2004) proposes a set of fault trees for a generic UAS. The generality of those fault trees makes them adaptable to many potential UAS platforms, but they require certain assumptions about the UAS (such as a gas-powered engine, two wings, separate ailerons and elevators, the presence of rudders, etc.) that are not applicable across UAS platforms. The FTA and FMEA analyses proposed for SUAS

platforms are normative rather than descriptive and are decidedly nonspecific because SUAS reliability data is not readily available for analysis (Dermentzoudis 2004).

Human Factors

The data available for large UAS mishaps tend to point to human factors as the most prevalent mishap factor. Different conclusions as to the extent and categories of human factors involved have been reached by researchers in part because there are a number of different ways to analyze the data resulting from mishap investigations. Due to the large number of classification schemes available, it is important to decide which one to use to classify risk factors prior to initiating analysis (Ballesteros 2007).

The DoD has developed the Department of Defense Human Factors Analysis and Classification System (DoD HFACS) to provide a common framework to classify and analyze human factors for mishap investigation (DoD 2005). This framework creates a taxonomy that is more descriptive than simply reporting “operator error” as a mishap cause (DoD 2005). The taxonomy is derived from work by Reason (1990) and Wiegmann and Shappell (2003) and is based on the concepts of active failures and latent failures/conditions resulting from hazards present in four different levels of responsibility. Mishaps are theorized to occur when hazards align across these four levels (see Figure 1). That is, it takes failures from the organizational and supervisory levels to permit the occurrence of preconditions for unsafe acts which ultimately result in active failures (mishaps). The DoD HFACS classification system has been used to categorize the human factors deemed responsible for large UAS mishaps. It relies on human

judgment to assign categories to the human error, so the conclusions resulting from analysis of these categorizations have varied by investigator, platform, and timeframe.

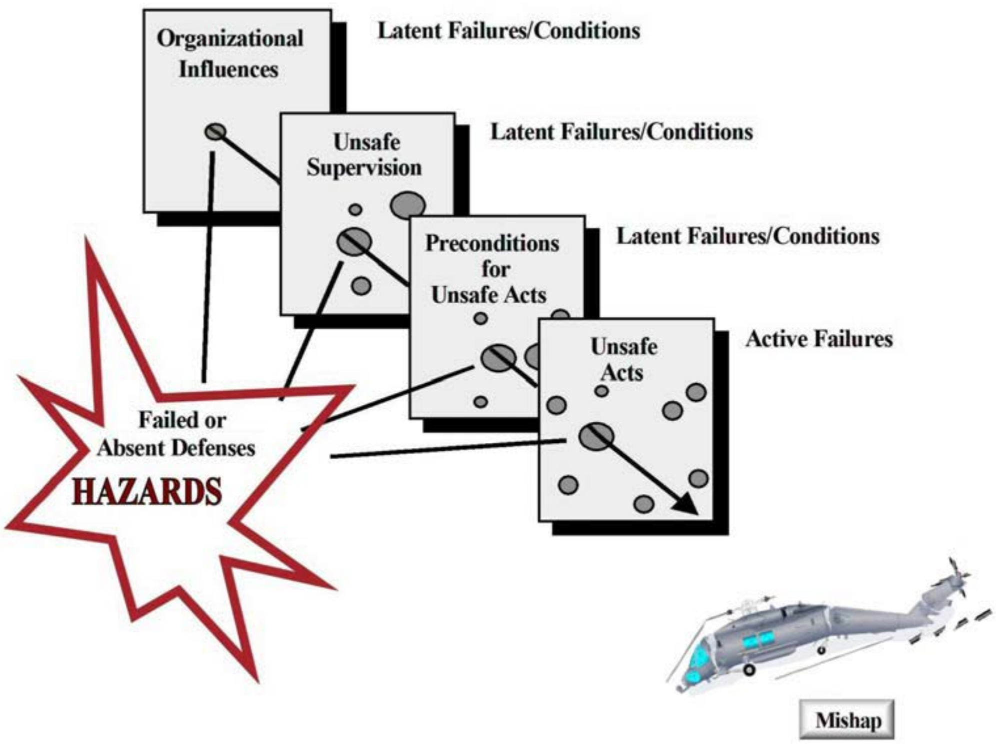


Figure 1. DoD HFACS levels (DoD 2005), based on work by (Reason 1990)

The major result from DoD HFACS analysis of aviation mishaps has been to identify Crew Resource Management (CRM) and Operational Risk Management (ORM) as main contributing factors to manned aviation mishaps, and Perceptual Errors as the main contributing human factor to Air Force UAS mishaps. An HFACS analysis of 124 Class A mishaps across manned aircraft revealed failures in CRM and ORM as common mishap causes (Gibb 2006). This meant that errors in communication between crewmembers, or failure to properly plan missions by ensuring aircrew proficiency, were

most often contributory to these mishaps. Large UAS, while generally having similar CRM and ORM considerations as manned aircraft, showed somewhat different results. Perceptual errors, suggestive of poor situational awareness, exacerbated by the peculiarities of UAS technology, were the leading cause of mishaps in US Air Force MQ-1 Predator UAS (Tvaryanas and Thompson 2006). Another analysis of the MQ-1 Predator with updated mishap data concluded that both perception and skill-based errors contributed the most to mishaps, but also shared similar latent failures (Tvaryanas and Thompson 2008). This means that MQ-1 mishaps that resulted from skill-based errors or perceptual factors had common antecedent hazards in the higher levels of the DoD HFACS taxonomy.

In a broad survey of UAS mishaps across all military branches, no major, common factors were isolated across the services (Tvaryanas, Thompson and Constable 2006). Instead, the Air Force tended to experience operator error from instrumentation/sensory feedback systems, automation and channelized attention, the Army saw latent organizational influences manifested as failures in guidance, training, and overconfidence, while the Navy and Marines were impacted by more complex factors closely associated with "workload and attention" and "risk management" (Tvaryanas and Thompson 2008). The HFACS analysis indicates that the Air Force has common failures in perceptual and sensory factors, possibly made worse by the technology employed by their UAS platforms. The other services and manned aircraft showed no common human factors. The commonality across Air Force mishaps gives hope that these latent and active human factors errors can be exploited for mishap rate improvement.

One caution should be noted about using human factors classification for mishaps; investigative biases may be present which are only reinforced by the labeling or relabeling of error (Dekker 2003). Researchers report the existence of hindsight bias, which is "the tendency for people with outcome knowledge to believe falsely that they would have predicted the reported outcome of an event" (Hawkins and Hastie 1990). This bias could impact the trustworthiness of mishap reports and the subsequent classifications of human error, as investigators may find fault in areas that are obvious in hindsight, but may not have been at the time of the mishap. This hindsight bias is "especially likely to occur when the focal event has well-defined alternative outcomes (e.g. win-lose)" (Hawkins and Hastie 1990), which makes it a potentially serious problem given the "mishap"-"no mishap" outcomes that are investigated. Hindsight bias, coupled with the practice of classifying error, "disembodies data…by excising performance fragments away from their context" (Dekker 2003). One theory of error is that humans perform erroneous actions which are viewed as rational from within their circumstances but which are not rational when viewed from the outside or in hindsight. Under this theory of "local rationality" any mishaps that occur are likely to reoccur as future individuals repeat the same locally rational acts, while a classification scheme on these errors merely provides a label to what in reality is a complex underlying problem (Dekker 2003). These underlying weaknesses in mishap reporting and classification are duly noted, but must be accepted in order to gain insights that can come from classification of SUAS mishaps, because these insights could lead to the mitigation of SUAS operational risks.

SUAS Risk Analysis

The risk scenarios commonly identified for UAS are mid-air collisions, ground impacts, and loss of the UAS platform. Despite many years of FAA data and several different models to predict the consequences of these risk scenarios, “there is currently no consensus on the specification of airworthiness regulations for UAS” (Clothier, et al. 2011). The major risks and their anticipated impacts are discussed in detail below.

The single most significant hazard for a UAS platform is a mid-air collision. This hazard is the primary one keeping civil UAS from being integrated into the NAS by the FAA (Clothier, et al. 2011). Mid-air collisions are a threat to both manned and unmanned aircraft operating in the vicinity of UAS. FAA reports through 2007 have only documented “a small number of incidents” of mid-air collisions between civil aircraft and remote control (R/C) airplanes, which all occurred between 1993 and 1998 and were attributed to lack of situational awareness in the manned aircraft, or violations of airspace rules and procedures by the remote pilots (Dalamagkidis, Valavanis and Piegl 2008). Despite the fact that no further data is available to quantify the consequences of mid-air collisions, (Dalamagkidis, Valavanis and Piegl 2008) believe that current regulations on R/C aircraft (which are vehicles similar to the size and performance of the SUAS under consideration in this thesis) are sufficient to ensure acceptable safety levels.

Ground impacts also pose a serious hazard for UAS operations. Several models have been developed to better quantify the risks associated with an impact to individuals and property. A blunt criterion estimation model for injury potential was developed for SUAS which computes the likelihood of a fatality based on a direct chest impact

(Magister 2010). When this model is applied to the airspeeds, frontal areas, and average mass of the SUAS considered in this thesis, most are shown to be at low risk for a fatality, even under the worst-case scenarios assumed by the model. A ground impact analysis performed by researchers at the Massachusetts Institute of Technology suggested that micro UAS (less than 2lb, less than 500ft altitude) posed a "relatively low risk" in general and that mini UAS (2 to 30 lb at 100 to 10,000ft altitude) could be flown over 95% of the country with low reliability requirements (Weibel and Hansman 2005). While the primary calculations of these models are in terms of fatalities, property on the ground can be damaged as well and injuries can be sustained, but neither of these two outcomes is taken as seriously as the potential for a fatality, and thus the numbers are not found in these types of analysis.

The last major risk scenario is the loss of the SUAS itself. This poses costs to the SUAS's organization both monetarily and in terms of lost mission capability. The minimum threshold for mishap reporting in the US Air Force is that of a Class C mishap, which involves any damage over $50,000. Many SUAS, like the ones flown by AFRL/RWWV, even if they were to be completely destroyed in a mishap, do not cost enough to meet that minimum threshold. When UAS mishaps occur, even if only resulting in minor damage to or loss of the UAS, they still have important policy and mission impacts. Four documented Canadian UAS mishaps in Afghanistan, while only damaging the aircraft themselves, nonetheless were said to have "created considerable risks for units that must retrieve these vehicles" and to have "increase[d] the workload on investigatory agencies" (Johnson 2008). Likewise, on the Eglin AFB range, there is a common UAS test requirement to report all aircraft that fly out of control or that exit

airspace boundaries. While any SUAS incidents that could meet these criteria may not have caused any harm to persons or property on the ground, the incidents still require reporting and possible investigation. The risk scenarios in SUAS operations are considerable, but the likelihoods of these scenarios occurring have not been investigated for SUAS.

Because little data exist on SUAS reliability, no empirical estimates are available to establish the likelihood of the aforementioned risk scenarios. Since risk assessment is comprised of a scenario, its likelihood of occurrence, and its consequence (Haimes 2009), the overall risks of SUAS operations have not been well-quantified. For example, the model for ground impact by Weibel and Hansman (2005) was used to calculate a necessary mean time between failures (MTBF) to ensure reliable UAS operation for a given population density, rather than computing actual reliability data from active systems. The model by Magister (2010) assumes a chest impact and merely quantifies the subsequent likelihood of a fatality, but does not seek to determine the probability of an SUAS colliding with a person's chest.

Overview of Mishap Prevention

The risks posed by UAS are deemed sufficient to warrant preventive actions. Many programs aimed at mishap reduction have been implemented for large UAS including: training, CRM, and medical screening. Additionally, research has examined pilot qualifications and the background experience necessary to make better UAS pilots. For the preventive actions that have had their effectiveness measured, the results are

largely inconclusive, demonstrating that there may not be a clear extension from them to prevent SUAS mishaps.

Many factors influence which prevention measures should be considered, including the cost and effectiveness of the proposed measures. Although literature on prevention is often found in a medical context, the basic principles of prevention are applicable across disciplines. The statement: “Research needs to be conducted before policies and programs are implemented when systematic reviews determine that scientific information is scant and where gaps in knowledge about prevention exist,” (Jones, Canham-Chervak and Sleet 2010) is as applicable to the medical field as it is to SUAS risk management. The health framework for prevention concludes that priority in preventive measures be allocated to those programs which have scientific evidence of effective prevention, and especially those which can produce it at the lowest cost (Jones, Canham-Chervak and Sleet 2010). This approach has been advocated in the aviation community as well: “in order to make best use of available resources prevention measures should focus on the areas with the greatest return…that are most manageable and those where the precursors are more susceptible to an antidote” (Gibb 2006). A survey of manned aircraft and large UAS preventive measures and their results may provide insight to determine the priority that decision makers should consider for SUAS mishap prevention.

Mishap Prevention Focused on Human Factors

One of the earliest manned aircraft mishap interventions was CRM, a training program introduced in the 1970s to reduce errors by focusing on human factors causes (Joint Aviation Authorities 2003). While CRM seems to produce positive responses in trainees, the gains from the program on flight safety are inconclusive (Salas, et al. 2001). Despite its lack of statistically significant success with manned aircraft, a CRM training program has been proposed as a preventive measure for the Indian Air Force's UAS operators (Sharma and Chakravarti 2005). CRM training has been introduced for USAF Predator operators (Nullmeyer, Herz and Montijo 2009), but the effects of that training have not been quantified. The use of CRM as an effective prevention for manned and unmanned aircraft mishaps remains to be seen, as insufficient data exist for analysis that may support or refute its efficacy.

Some preventive measures for large UAS have focused on pilot qualifications and screening. Given that human factors play a significant role in causing UAS mishaps, studies have been undertaken to determine if proper pilot selection can prevent mishaps. Schreiber, et al. (2002) found that on a high-fidelity Predator flight simulator, about 150-200 hours of previous flight experience was required to match the performance of Air Force pilots that are currently selected for Predator training. This means that an individual with a civilian pilot's license or one who had just completed T-38 training was as skilled at the simulation as an operational pilot with no previous Predator experience, implying that the skills needed for UAS operation may be enhanced with any prior flight experience. The study's authors are quick to note that experienced manned pilots who

switch over to UAS may have to “unlearn” some skills they have learned in the cockpit as the sensory environment is much different for UAS (Schreiber, et al. 2002). In a study by Tvaryanas, Thompson and Constable (2006), which looked at multiple UAS platforms across the services, the authors found that “experienced military pilot UAV operators made as many bad decisions as enlisted UAV operators without prior military flight training or experience” which suggests that limiting UAS pilots to rated officers may not improve overall flight safety. Lastly, the FAA has proposed screening UAS pilots for civil operations with a second-class medical certification in the hopes of reducing the level of risk associated with pilot incapacitation (Williams 2007). This recommended certification level is justified by noting that manned aircraft with similar missions that operate in the proposed airspace have second-class certification requirements for their pilots, although it is conceded that waivers are available for anyone who can demonstrate safe aircraft operation (Williams 2007). Since these are proposed rules, no data exist to quantify their effect on flight safety. Pilot screening and minimum qualification requirements for UAS operations may only be beneficial when prior flight experience is taken into account, regardless of rank or medical status.

Mishap Prevention Focused on Technical Factors

Technical preventive measures are introduced frequently in the UAS world: this thesis itself is based on using data gathered while testing new technical innovations for SUAS platforms. The technical risk factors for UAS mishaps previously discussed are largely inconclusive and may not justify a technical intervention. While it is assumed that

technological advances will proceed as the development of UAS platforms proceeds, several authors caution that adding technology to already complex systems may degrade performance. “There will be situations where the solution increases the complexity of the system and, as a secondary effect, reduces the risk of one factor while increasing that of another” (Ballesteros 2007). These effects are most pronounced in systems with “interactive complexity” and “tight coupling”, which refers to systems like aircraft where cause and effect are nonlinear with quick propagation of events through the system (Perrow 1999). Fixes to these systems, “including safety devices, sometimes create new accidents” (Perrow 1999). For that reason, technological fixes should be approached cautiously lest their added complexity increase the risk of the type of accidents they seek to prevent.

Several specific preventive technical measures have been proposed to increase the reliability and safety of UAS operations, primarily automated landing capability and sense-and-avoid. These two measures are proposed to allow UAS to perform at levels of safety equivalent to manned aircraft. This is an important consideration for integrating UAS in the NAS (Mejias, et al. 2009), and has potential to improve reliability figures for UAS across all operational domains.

Automated landing capabilities are cited as having great potential to reduce UAS mishaps. The RQ-7 Shadow UAS, flown by the Army, is equipped with a tactical automated landing system (TALS) to eliminate external pilot landing errors. TALS is far from perfect, causing 25% of Shadow mishaps as analyzed by (Williams 2004). This system also requires operators to setup a landing site in advance with equipment preplaced near the runway. In research conducted by (Mejias, et al. 2009), an automated

landing system was proposed that allows logic onboard the UAS to select the optimal landing site in an emergency situation, eliminating the need for ground crew and setup time. Regarding increases in automation in general, (Williams 2004) states that “the use of automation to overcome human frailties does not completely solve the problem, as the automation itself can fail”. These automated landing approaches have promise for reducing UAS mishaps, although affirmative results have not yet been obtained and their added complexity may be problematic.

Sense and avoid (SAA) is a preventive measure that would allow UAS to detect other airborne traffic and avoid a collision. This technology has been mandated by regulations, particularly FAA Order 7610.4, which requires SAA systems to perform as well as manned aircraft (Carney, Walker and Corke 2006). SAA would lower the probability of the most severe risk scenario facing UAS (a mid-air collision), and is a requirement before UAS can be integrated into the NAS. These systems have not yet been implemented on UAS platforms despite some successful demonstrations, because “testing without access to the NAS is problematic” (Dalamagkidis, Valavanis and Piegl 2008).

AFRL’s SUAS Program Background

The Air Force Research Laboratory’s Munitions Directorate (AFRL/RW) has been performing flight experiments on SUAS since at least 2005. The directorate uses computer aided design and manufacturing techniques with rapid-prototyping equipment to create and modify SUAS vehicles for a variety of missions. AFRL/RW has produced

several vehicles of note, including the BATCAM and GENMAV. The Flight Vehicles Integration Branch (AFRL/RWWV) not only designs aircraft but tailors existing commercial-off-the-shelf (COTS) remote control aircraft for flight experiments. AFRL/RWWV has a varied mission, both designing and flying experimental SUAS to determine the feasibility of new technologies, and integrating customer payloads into existing SUAS platforms to provide flight data.

The BATCAM (see Figure 2) is an example of an aircraft designed by AFRL/RW to push the technological boundaries. The BATCAM was designed as a battlefield surveillance platform for the USAF's Battlefield Air Operations (BAO) kit. The vehicle is a man-portable SUAS capable of being hand-launched by operators and was designed to prove that compact surveillance vehicles were technologically feasible.

Figure 2. BATCAM SUAS developed by AFRL (Abate, Stewart and Babcock 2009)

The GENMAV (see Figure 3) is an aircraft designed by AFRL/RW as a technology demonstration platform. The GENMAV was originally conceived as a

baseline configuration for basic aerodynamic research. It has been used for that research, but has also been modified to characterize flight maneuvers with flexible wings and has been outfitted with different payloads for parachute recovery experimentation. These are two of the over two-dozen SUAS flown by AFRL/RWWV since 2005.

Figure 3. GENMAV SUAS developed by AFRL (Abate, Stewart and Babcock 2009)

The aircraft flown by AFRL span a wide range of the SUAS category. They vary in wingspan from 20 inches to 11 feet, with takeoff weights under 100 pounds. The larger SUAS are gasoline powered while the smaller ones are battery-powered with electric motors. Most are equipped with miniaturized autopilot technology to enable semi-autonomous flight. The SUAS have three flight modes: autonomous flight with waypoints preloaded into memory, semi-autonomous flight where the pilot provides directional inputs while the autopilot maintains altitude and speed, and manual flight where all commands are given by the pilot. Some aircraft are flown exclusively in

autonomous mode, others are flown exclusively manually, and the rest are flown with a mix of both depending on their missions and the goals of that particular flight experiment.

AFRL/RWWV operates under AFRL Instructions for its flight test program. Two primary documents govern its SUAS operations: AFRLI 61-103 “AFRL Research Test Management” and AFRLMAN 99-103 “AFRL Flight Test and Evaluation”. The first document outlines the general policy for testing in AFRL. It contains a risk assessment matrix (see Figure 4) for test planning which allows program managers to determine the level of risk each test poses which in turn determines the appropriate level of approval. The dearth of SUAS data makes filling out this risk matrix highly subjective, as the consequences are frequently unknown and their likelihoods have not been formally quantified. AFRLI 61-103 defines a mishap as “unplanned events or range operations resulting in loss/damage to DoD or private property, injury, departure from range boundaries, or public endangerment”. The second document, AFRLMAN 99-103, defines Class A through C mishaps much like the DoD classification, except that the dollar figures are lower (AFRLMAN 99-103 is an older document). The AFRL manual notes that Class D mishaps are not applicable to flight-related mishaps and adds a Class E category:

> Class E Events: These occurrences do not meet reportable mishap classification criteria, but are deemed important to investigate/report for mishap prevention. Class E reports provide an expeditious way to disseminate valuable mishap prevention information.

HAZARD SEVERITY CATEGORY				
HAZARD PROBABILITY	*Catastrophic-I* Could result in death, permanent total disability or system/ facility loss>$1M	*Critical-II* Could result in permanent partial disability, injuries or illness that may result in hospitalization of >3 personnel, or system/ facility loss >$200K but <$1M.	*Marginal-III* Could result in injury or illness resulting in >1 day of lost work, system/facility loss >$20K but <$200K.	*Negligible- IV* Could result in injury or illness not resulting in lost work time, system /facility loss >$2K but <$20K.
FREQUENT-A Likely to occur often in the life of an item or during an event.	**1**	**3**	**7**	**13**
PROBABLE-B Will occur several times during the life of an item or during an event.	**2**	**5**	**9**	**16**
OCCASIONAL-C Likely to occur some time in the life of an item or during an event	**4**	**6**	**11**	**18**
REMOTE-D Unlikely, but possible to occur in the life of an item or during an event	**8**	**10**	**14**	**19**
IMPROBABLE-E Highly unlikely to occur in the life of an item or during an event	**12**	**15**	**17**	**20**

Figure 4. Risk Assessment Matrix for AFRL Testing. Boxes 1 – 4 denote High Risk tests, 5 – 9 are Medium Risk tests, and 10 – 20 are Low Risk tests (AFRLI 61-103)

Most of the aircraft flown by AFRL/RWWV do not meet the minimum cost levels required for Class C mishap reporting. That is, if an SUAS were to crash and be completely destroyed, it would not have caused enough damage (in dollars) to warrant a Class C mishap investigation and report. Likewise, AFRL/RWWV's SUAS fleet is composed of mostly small vehicles that are highly unlikely to cause fatalities even under worst-case scenarios. Therefore, the term "mishap" as defined by the DoD is not applicable to the majority of AFRL's SUAS. Instead, the term "failure" is used for the remainder of this thesis. An SUAS failure is said to occur in AFRL/RWWV's flight experimentation program whenever required flight experiment data is not obtained due to an SUAS or SUAS operator fault.

Since the end product of the flight experiments are data, any SUAS action that prevents the planned data from being collected is a failure. For example, if an SUAS fails to cleanly launch and crashes on takeoff, that is deemed a failure, as the data from that flight is lost. If an SUAS loses communication with the ground station and is forced to land before all test points are completed, that, too is a failure, even though no damage occurred to the platform. If an SUAS flies approach too steeply and breaks its landing gear after all test points have been completed, that is not considered a failure, despite the occurrence of damage. The term “failure” is an objective measure of the SUAS’s ability to execute its mission for AFRL and is distinct from “damage”, which is quantified monetarily to determine a mishap category.

Logistic Regression Modeling

Logistic regression is an analytical technique used to construct a model describing the relationship between a dependent variable with a discrete response and one or more explanatory variables (Hosmer and Lemeshow 1989). Dichotomous responses (using “0” or “1” to indicate the nonoccurrence or occurrence of some outcome, respectively, for example) violate many of the assumptions of Ordinary Least Squares (OLS) regression including homoscedasticity and normality of residuals (Menard 2002). Additionally, OLS regression will produce a model whose range is $-\infty$ to $+\infty$, which violates the 0 to 1 range for a binary discrete response. An example of a dichotomous response is shown in Figure 5. An OLS regression on the data in this plot would be less than 0 for low values

of the explanatory variable, and would exceed 1 for high values of the explanatory variable.

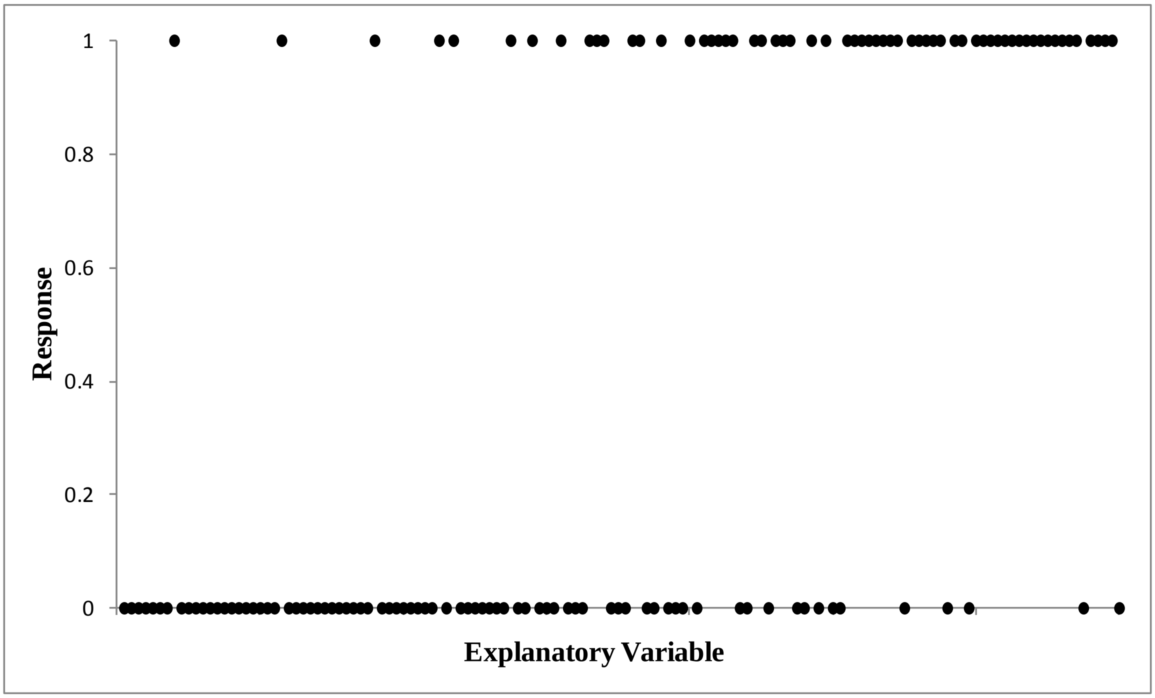


Figure 5. Example plot of a dichotomous response.

Logistic regression addresses these issues by producing a model with a continuous range from 0 to 1 that indicates the probability of membership in group 1 given the values of explanatory variables (Menard 2002). Logistic regression models also have the interpretive benefit of fitting the rate of occurrence of the response variable. Figure 6 is a logistic regression model fit to the data from Figure 5 when the explanatory variable is divided into seven equally sized groups and the corresponding response rate is modeled against their respective midpoints.

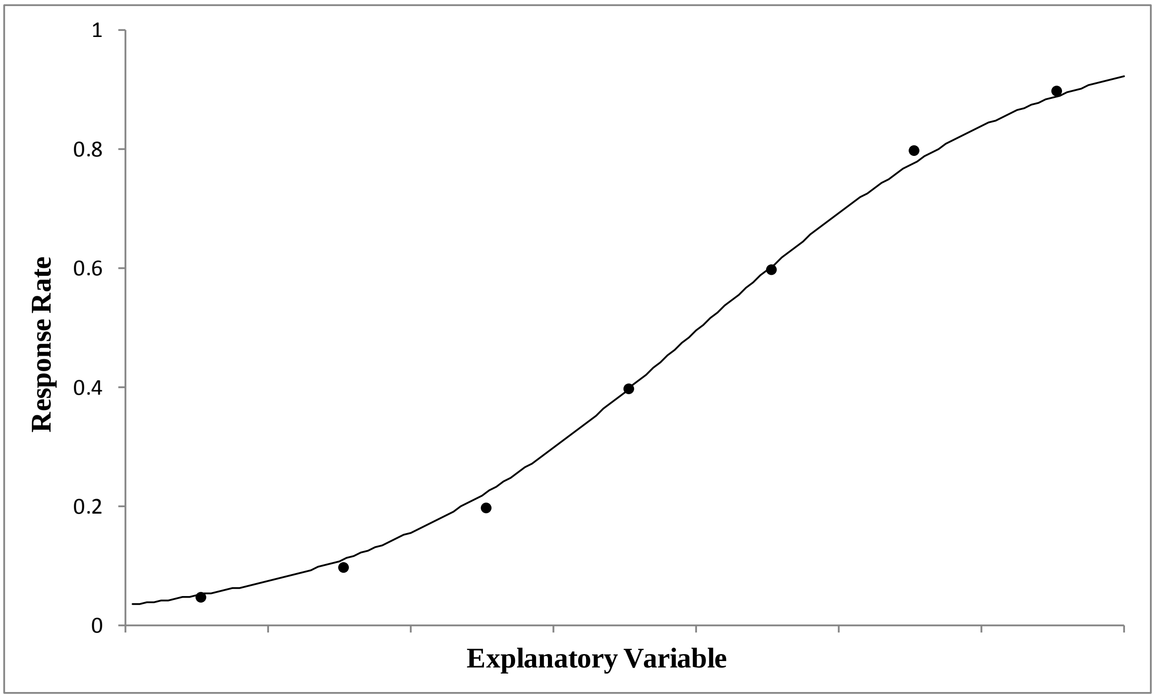


Figure 6. Logistic regression model, fitted to rate data from Figure 5.

In general for a logistic regression model, let y be the dependent variable and $\vec{x}$ be the vector of explanatory variable values. The probability of interest is expressed as:

$$\Pr\{y = 1|\vec{x}\} = \pi(\vec{x}).$$

The logistic distribution is used to model the probability. It takes the form:

$$\pi(\vec{x}) = \frac{e^{g(\vec{x})}}{1 + e^{g(\vec{x})}}$$

where $g(\vec{x})$ is known as the logit transformation and can be expressed as:

$$g(\vec{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = ln\left[\frac{\pi(\vec{x})}{1 - \pi(\vec{x})}\right].$$

The logit transformation is comparable to functions used in OLS regression because $g(\vec{x})$ is continuous with a range from $-\infty$ to $+\infty$ and is linear in its parameters. The logistic distribution is restricted to a 0 to 1 continuous range, it is a flexible function,

and it lends itself well to interpretation(Hosmer and Lemeshow 1989). The parameters of the logistic function, β_i, are usually estimated iteratively using maximum likelihood methods and are important for the model's interpretation.

Given a probability of an event occurring, $\pi(\vec{x})$, the odds of that event occurring are:

$$\frac{\pi(\vec{x})}{1-\pi(\vec{x})} = e^{g(\vec{x})}.$$

The exponentiation of any parameter β_i represents a ratio of odds when the explanatory variable x_i is increased by one unit. To see this, consider two logistic distributions, $\pi_1(\vec{x})$ and $\pi_2(\vec{x})$. Let $g_1(\vec{x})$ and $g_2(\vec{x})$ be the logit functions associated with each of these distributions, respectively, where $g_1(\vec{x})$ is identical to $g_2(\vec{x})$ except that variable x_i has been increased by one unit:

$$g_1(\vec{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i(x_i + 1) + \cdots + \beta_p x_p$$

and

$$g_2(\vec{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i(x_i) + \cdots + \beta_p x_p.$$

The odds ratio of these two logistic distributions becomes:

$$\frac{\frac{\pi_1(\vec{x})}{1-\pi_1(\vec{x})}}{\frac{\pi_2(\vec{x})}{1-\pi_2(\vec{x})}} = \frac{e^{g_1(\vec{x})}}{e^{g_2(\vec{x})}}$$

$$= \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_i(x_i+1)+\cdots+\beta_p x_p}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_i(x_i)+\cdots+\beta_p x_p}}$$

$$= \frac{e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_i(x_i+1)} \ldots e^{\beta_p x_p}}{e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_i x_i} \ldots e^{\beta_p x_p}}$$

$$= \frac{e^{\beta_i(x_i+1)}}{e^{\beta_i x_i}}$$

$$= \frac{e^{\beta_i x_i+\beta_i}}{e^{\beta_i x_i}}$$

$$= \frac{e^{\beta_i x_i} e^{\beta_i}}{e^{\beta_i x_i}}$$

$$= e^{\beta_i}.$$

While a parameter in OLS regression reflects the change in the mean response variable due to an increase in one unit of the explanatory variable, the parameters in logistic regression represent the natural logarithm of the change to the odds ratio of the response.

When building a logistic regression model, a stepwise strategy is often employed with the maximum p-value of entry into the model, p_E, set to a value between 0.15 and 0.20, although this may be relaxed to $p_E = 0.25$ if the analyst desires to include a greater number of potential explanatory variables (Hosmer and Lemeshow 1989). The minimum p-value of removal from the model, p_R, should be set slightly larger than p_E, with typical values being $p_E = 0.15$ and $p_R = 0.20$. Terms in the model are assumed linear in the logit, an assumption tested using the Box-Tidwell transform, which tests for the significance of the coefficient β_i on the new term $x_i \ln x_i$ when it is added to the model. A significant coefficient (usually at the $\alpha = 0.05$ level) means there is nonlinearity in the logit (Hosmer and Lemeshow 1989). Likewise, interactions should be assessed among variables where different response rates are expected at different levels.

To assess the model's classification accuracy a confusion matrix (also called a classification table) can be used, which shows the counts of true positives, true negatives, false positives, and false negatives obtained for a specified probability cutoff, usually $\pi_0 = 0.5$. A more informative assessment is found by using a receiver operating characteristic (ROC) curve (Agresti 2002). This curve plots the sensitivity of the model as a function of (1- specificity) for the range of π_0. The higher the area under the curve, the better the model is at classification, with 0.5 indicating that a model classifies no better than random guessing (Agresti 2002).

Artificial Neural Networks

An Artificial Neural Network (ANN) is an information processing system that can be used for classification or regression analysis (Steppe 1994) and (Bauer 2011). For classification networks, an input vector's information is extracted by the network and processed in parallel by a number of "neurons" or nodes, which produce a classification output. The input vector is a collection of the values of all independent variables (known as "features" in the neural network) for a single instance. In the case of SUAS failure data, an input vector would consist of the values of all the features deemed important to the model for one flight. The model processes one input vector per flight and compares its classification of "Mishap" or "No Mishap" to the known flight outcome, which is supplied with the input vector.

A typical feedforward network takes the input vector's values and processes them forward through the "hidden layer" of nodes to the output layer, which yields a

classification. It is called a “feedforward” network because the information travels forward and is never fed back to any previous nodes. A simple feedforward artificial neural network for classification with one hidden layer and two classifier nodes (which is the neural network structure used herein for SUAS failure analysis) is shown in Figure 7.

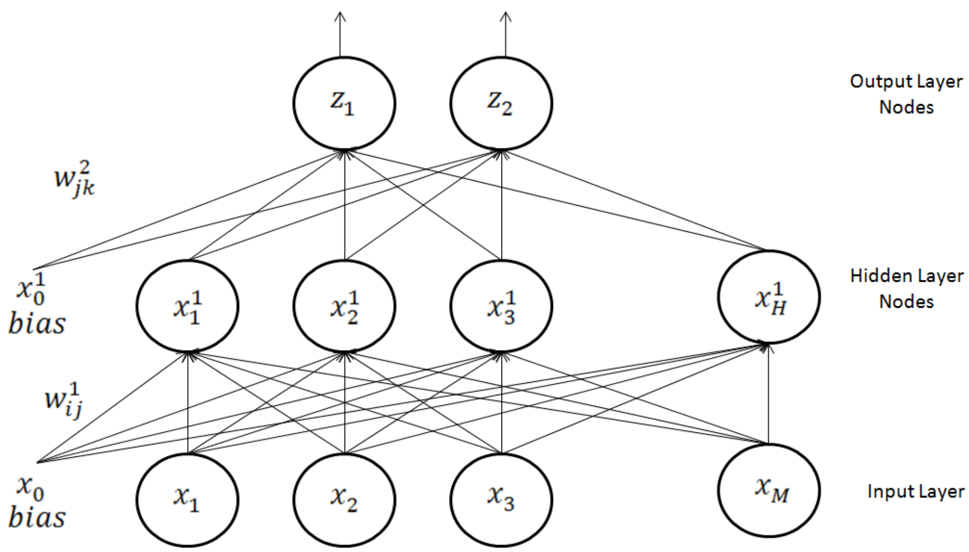


Figure 7. Feedforward Neural Network structure with one hidden layer and two output nodes for classification. Based on a diagram from (Steppe 1994).

To process the data, each feature’s input value (the windspeed, number of total flights, or days since pilot’s last flight, for example) is first normalized by subtracting that feature’s mean and dividing by its standard deviation (Bauer 2011). This normalized input is multiplied by a unique numerical weight (w_{ij}^1 in Figure 7) before it enters each hidden layer node (x_n^1 in Figure 7). Within each node in the hidden layer, the weighted inputs of all features are summed and then standardized to a 0 to 1 range using a squashing function, such as the sigmoidal activation function (Steppe 1994). The sigmoid

function takes an input and transforms it to a 0 to 1 range. For a given numerical input, x, it takes the form:

$$\frac{1}{1+e^{-x}}.$$

This is equivalent to the logistic distribution, only with x expressed as a negative exponent in the denominator rather than as a positive exponent in both the numerator and denominator:

$$\frac{1}{1+e^{-x}} \times \left(\frac{e^x}{e^x}\right) = \frac{e^x}{e^x + e^{-x+x}} = \frac{e^x}{1+e^x}.$$

This function ensures that any numerical input is restricted to 0 to 1 output; hence it is referred to as a "squashing" function.

After the weighted, summed values are squashed to the 0 to 1 range by the sigmoid function, each hidden layer node's output is then fed forward to be multiplied by a numerical weight (with weights w_{jk}^2 from Figure 7). All of these squashed, weighted values from the hidden layer of nodes then become the inputs for the output layer of nodes. The output nodes sum these inputs and squash them exactly as the hidden layer nodes previously did. Each output node corresponds to a possible outcome. The node with the highest output value gives the input vector its group classification. For the SUAS data, the flight is classified in group 1 ("Mishap" or "Damage") if output node 1 produces a value larger than output node 0. If output node 0 produces the larger of the two values, the flight is classified as group 0 ("No Mishap" or "No Damage"). To provide better insight into the neural network process, the inner workings of a hidden layer node (x_2^1) are depicted in Figure 8. Some possible features (SUAS flight variables which have been

normalized) are shown in the input layer for explanatory purposes, but do not necessarily reflect the significant features of the final model.

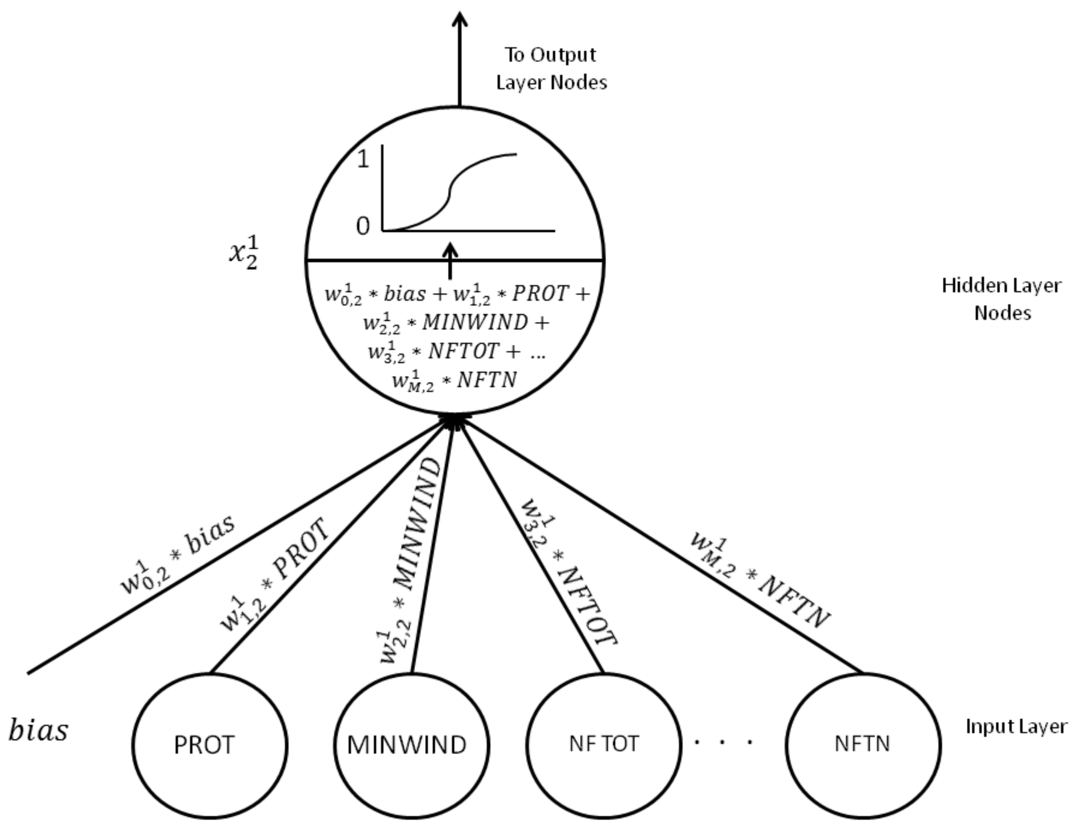


Figure 8. A hidden layer node in a hypothetical feedforward network.

Artificial neural networks improve their performance by using learning algorithms. These algorithms allow the neural network to adjust its weights according to a known classification for the given input. The network is trained to minimize error between its output and the truth data provided by the user. The learning algorithm used here for the SUAS failure data is called backpropagation. This algorithm minimizes the mean squared mapping error by updating both levels of weights after each input vector is

fed through the network by performing a gradient search of the error surface (Bauer 2011). Essentially, the network compares its numerical output with the actual “0” or “1” flight outcome. It then updates its weights to provide the most dramatic decrease in the squared difference between the network’s output and the actual output. After the input data associated with each flight is fed forward, the network “learns” the best adjustment of its weights to provide more accurate results.

The network is “trained” with only a subset of the data (usually 60-70%) while the remaining data are partitioned for validation and testing. Backpropagation is used to adjust the network’s weights for the training subset of data only. The validation data are fed forward through the network to determine their mapping error. In general, the network is considered optimized when the validation data error is at a minimum. Since a neural network with enough nodes can map an arbitrarily complex surface, the validation data set is used to prevent overfitting. Overfitting occurs when the network learns the training data so well that it no longer generalizes to other, similarly collected data (which is what the validation data represents). Once the network is optimized, the test data is used as an independent check of the overall classification accuracy of the network.

As with logistic regression, determining which input features are salient to the model is important for parsimony and interpretation. Two primary saliency measures have been proposed for neural network features: weight-based saliency measures and derivative-based saliency measures (Bauer 2011). Weight-based measures take the sum of the squares of the lower-level of weights ($w^1_{i,j}$ in Figure 7) for a given feature under the assumption that the more salient features have weights significantly greater or less than 0 whereas less salient features will tend to have weights of a smaller magnitude

(Tarr 1991). Derivative-based saliency measures compute partial derivatives of the network's output with respect to feature inputs to determine a saliency measure (Bauer 2011). In both cases, the saliency of a candidate feature (considered for removal from the model) can be compared to an injected noise feature, which is usually a uniform random variate from 0 to 1 (Bauer 2011). If the candidate feature differs in a statistically significant manner from the noise, it can be considered salient to the model.

The signal-to-noise ratio (SNR) saliency measure proposed by Bauer, Alsing and Greene (2000) is used for SUAS failure modeling. This measure is weight-based and uses the injected noise input as a comparison for all candidate features. The saliency measure is computed by taking the ratio of the sum of squares of the weights for the candidate feature *i* and the injected noise *n* and converting to a decibel scale (Bauer, Alsing and Greene 2000):

$$SNR_i = 10 \log_{base10} \frac{\sum_{j=1}^{J} \left(w_{i,j}^{1}\right)^2}{\sum_{j=1}^{J} \left(w_{n,j}^{1}\right)^2}.$$

Neural networks are randomly initialized, a fact which can often produce different results for the same inputs. To account for this randomness, the SNR saliency measure is computed for each feature for some number of neural networks (usually between N = 10 and N = 30). The measure can be used to rank order the features, after which the least significant feature (lowest ranked) is removed and the average classification accuracy of the retrained networks is computed (Bauer, Alsing and Greene 2000). When there is a significant drop-off in the classification accuracy after a feature is removed, the last feature removed is retained in the network. When there is not a clear drop off, the analyst or decision maker uses their discretion to determine the cut-off point at which the

classification accuracy is acceptable. The remaining features are considered significant in the model. As with logistic regression, confusion matrices and ROC curves are used to assess the classification performance of the networks.

Summary of Literature Review

The risks to SUAS are numerous. Prior experience suggests that if SUAS are comparable to their larger unmanned counterparts, they are at greatest risk of a failure from human error. Factors expected to reduce this risk are pilot experience, pilot currency, and any prior, manned flight experience. Since only one of AFRL/RWWV's SUAS pilots held an FAA-certified pilot's license, and he flew for 8 flights (less than 1% of total flights), only pilot experience and currency are investigated in this thesis. Currency is measured as days since a pilot's last flight.

The next most likely source of risk is weather. Temperature is not expected to affect pilot performance, whereas wind speed has great potential to contribute to SUAS failures. Both ambient temperature and surface wind speeds are investigated for their contributions to SUAS failures, as well as experience at given flight locations, which may exhibit unique local weather patterns.

The generic catchall factor of organizational experience suggests that failure rates will decrease with greater experience. Total organizational number of flights are investigated as a factor for its impact on failure rates. Additionally, the number of flights on specific air frame types ("BATCAM" or "GENMAV", for example) are investigated to determine if failure rates decrease with specific platform experience. Number of flights

on a given tail number (“BATCAM #12” or “GENMAV #3”, for example) are also investigated to determine its relationship to failure rates.

Although not mentioned in any research above, interval values are investigated to determine if the time between flights (for air frame, tail number, autopilot type, mission, pilot and location) affects the failure rate. Lastly, since research indicated that different types of aircraft experienced unique failure modes and rates, the data are analyzed while controlling for type of SUAS, whether an AFRL-designed prototype, or a COTS air frame. Likewise, the data are controlled for whether or not the SUAS was flown manually or assisted by autopilot, as these different modes of flight are likely to affect failure rates and types. These control factors are included when they are found to be statistically significant to the model, and are disregarded if they are not.

III. Methodology

Overview of Dataset and Modeling Approach

The dataset used for this thesis was derived from all available flight test reports (n = 854) from AFRL/RWWV over the years 2005-2009. The dataset consists of 20 explanatory variables and three outcome variables whose values were extracted from the text or context of the flight reports (see Table 1). Not every flight has complete data: for example, some are missing wind speeds and temperatures while others (particularly those not flown on the Eglin range) are missing flight failure or damage outcomes. Every flight was entered into the database so that interval values could be determined (for example, if there is no data for failure or damage for tail number 12 when it last flew, the number of days between flights is still recorded on its next flight and its total number of flights is incremented).

When dealing with missing data values, there are a few remedies that may be adopted. If the data that are missing meet certain randomness and ignorability assumptions, there are maximum likelihood estimation and imputation techniques that can maximize the available data by replacing these missing values while minimizing any bias introduced (Allison 2009). The technique adopted here is listwise deletion, in which a flight is deleted from the model if it is missing a value in a variable considered important to that model. This technique discards much data, but is “honest” in that it usually results in large but accurate standard error estimates, which some other techniques may artificially lower (Allison 2009).

Table 1. Code listing for all variables.

Code	Description	n	Min	Max	Mean
DSAFLF	Days Since Air Frame Last Flew	825	0	911	8.28
DSAPTLF	Days Since Autopilot Type Last Flew	766	0	486	5.43
DSLF	Days Since Last Flight	853	0	36	2.08
DSLM	Days Since Last Mission	853	0	36	7.71
DSLOCLF	Days Since Location Last Used	796	0	729	8.37
DSPLF	Days Since Pilot Last Flew	787	0	484	7.22
DSTNLF	Days Since Tail Number Last Flew	668	0	308	11.6
MAN	(0 = Autopilot, 1 = Manual)	854	0	1	0.0842
MAXWIND	Maximum Forecast Wind (kts)	738	0	25	8.16
MINWIND	Minimum Forecast Wind (kts)	738	0	15	4.56
NFAF	Number of Flights on Air Frame	854	1	251	59.2
NFAPT	Number of Flights on Autopilot Type	772	1	447	146
NFLOC	Number of Flights at Location	805	1	564	206
NFP	Number of Flights by Pilot	805	1	481	162
NFTN	Number of Flights on Tail Number	771	1	42	10.2
NFTOT	Number of Flights Total	854	1	854	428
PROT	(0 = COTS Aircraft, 1 = Prototype)	854	0	1	0.712
TEMP	Forecast Ambient Temperature (F)	751	25	95	71.9
TIME	Time of Day Mission Started (ex: 0800 = 8.0, 1545 = 15.75)	704	4.5	22	9.96
WINDDIFF	(MAXWIND - MINWIND)	738	0	20	3.60
DAMAGE	(0 = No SUAS Damage, 1 = Damage)	751*	0	1	0.233
FAILURE	(0 = No Failure, 1 = Failure)	754**	0	1	0.385
FAILURE3	(0 = Human Error, 1 = Mechanical Error, 2 = No Failure)	754**	0	2	1.42

* There are n = 542 flights with complete records and DAMAGE outcomes
**There are n = 540 flights with complete records and FAILURE outcomes

A series of logistic regression models were constructed to assess the significance of measured variables on different outcomes associated with SUAS failures. All models were built with JMP 9.0 software using a forward stepwise algorithm with $p_E = 0.15$ and $p_R = 0.20$.

Logistic Regression Failure Prediction Model

For the Logistic Regression Failure Prediction Model, the response variable under consideration is FAILURE (0 = no failure, 1 = failure). The parameter estimates for the resulting model (n = 672, with prior probabilities Pr{FAILURE = 0} = 0.64 and Pr{FAILURE = 1} = 0.36) are shown in Table 2.

Table 2. Parameter estimates and significance for the Logistic Regression Failure Prediction Model

Term	Estimate	Lower 95%	Upper 95%	Std Error	Chi Square	Prob>ChiSq
Intercept	-2.67096	-3.76805	-1.57386	0.55974	6.181	0.013
PROT	1.81659	1.29893	2.33425	0.26411	47.309	0.000
MAN	0.74216	0.03315	1.45117	0.36174	4.209	0.040
NFTOT	-0.00101	-0.00187	-0.00015	0.00044	5.297	0.021
NFAF	-0.00533	-0.00854	-0.00212	0.00164	10.597	0.001
NFTN	0.03781	0.01698	0.05863	0.01063	12.661	0.000
TEMP	0.01379	0.00084	0.02675	0.00661	4.353	0.037

The corresponding odds ratios for a one unit increase in each explanatory variable are given in Table 3.

Table 3. Odds ratios for a one-unit increase for variables in the Logistic Regression Failure Prediction Model

Term	Odds Ratio	Lower 95%	Upper 95%
PROT	6.15083	3.66538	10.32161
MAN	2.10047	1.03370	4.26815
NFTOT	0.99899	0.99813	0.99985
NFAF	0.99468	0.99150	0.99788
NFTN	1.03853	1.01713	1.06039
TEMP	1.01389	1.00084	1.02711

The model's variables did not exhibit significant nonlinearity in the logit with Box-Tidwell terms incorporated in the model, nor were any significant interactions found. The ROC curve and confusion matrix are shown in Figure 9 and Figure 10 respectively. The area under the curve (AUC) is 0.718, with a 69.6% hit rate for classification.

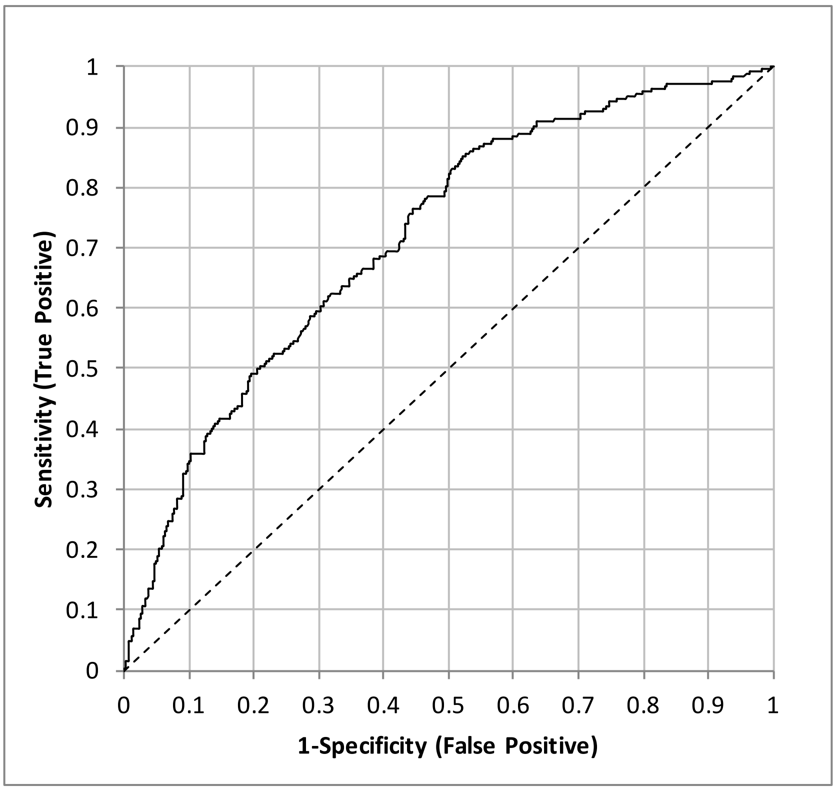


Figure 9. ROC Curve for Logistic Regression Failure Prediction Model. AUC = 0.718.

		Predicted	
		0	**1**
Actual	**0**	368	62
	1	142	100

Figure 10. Confusion Matrix for Logistic Regression Failure Prediction Model. Hit Rate = 69.6%.

Logistic Regression Damage Prediction Model

For the Logistic Regression Damage Prediction Model, the response variable under consideration is DAMAGE (0 = no SUAS damage, 1 = damage to SUAS). The initial model (n = 678, with prior probabilities Pr{DAMAGE = 0} = 0.783 and Pr{DAMAGE = 1} = 0.217) exhibited nonlinearity in the logit due to the NFTN variable. The Box-Tidwell transform is left in the model to correct the nonlinearity. The parameter estimates and significance are shown in Table 4. The corresponding odds ratios for a one unit increase in each explanatory variable (except NFTN) are given in Table 5.

Table 4. Parameter estimates and significance for the Logistic Regression Damage Prediction Model.

Term	Estimate	Lower 95%	Upper 95%	Std Error	Chi Square	Prob>ChiSq
Intercept	-1.61308	-2.16840	-1.05776	0.28333	4.60760	0.032
PROT	1.42970	0.90387	1.95553	0.26828	28.40087	0.000
NFAF	-0.00842	-0.01259	-0.00425	0.00213	15.66932	0.000
MAN	0.58012	-0.13653	1.29677	0.36564	2.51739	0.113
NFTN	-0.18407	-0.35776	-0.01039	0.08862	4.31479	0.038
NFTN*ln(NFTN)	0.05637	0.00902	0.10372	0.02416	5.44358	0.020

Table 5. Odds ratios for a one-unit increase for variables in the Logistic Regression Damage Prediction Model.

Term	**Odds Ratio**	**Lower 95%**	**Upper 95%**
PROT	4.17746	2.46914	7.06766
NFAF	0.99161	0.98749	0.99576
MAN	1.78625	0.87238	3.65748
NFTN	-	-	-
NFTN*ln(NFTN)	-	-	-

The odds ratio for NFTN cannot be directly obtained by exponentiating its parameter because the Box-Tidwell transformed term, which is a function of NFTN, affects the model's predicted probability. The odds ratio for NFTN is not constant, but is a function of its present value. The derivation of the odds ratio for a one-unit increases in NFTN is shown below. In general, the odds ratio can be expressed as:

$$OR_{NFTN} = \frac{Odds\ with\ (NFTN+1)}{Odds\ with\ NFTN}$$

$$= \frac{e^{Intercept+\beta_{PROT}PROT+\beta_{NFAF}NFAF+\beta_{MAN}MAN+\ \beta_{NFTN}(NFTN+1)+\beta_{NFTN*\ln(NFTN)}(NFTN+1)*\ln\,(NFTN+1)}}{e^{Intercept+\beta_{PROT}PROT+\beta_{NFAF}NFAF+\beta_{MAN}MAN+\ \beta_{NFTN}(NFTN)+\beta_{NFTN*\ln\,(NFTN)}NFTN*\ln\,(NFTN)}}$$

$$= \frac{e^{\beta_{NFTN}(NFTN+1)+\beta_{NFTN*\ln(NFTN)}(NFTN+1)*\ln\,(NFTN+1)}}{e^{\ \beta_{NFTN}(NFTN)+\beta_{NFTN*\ln\,(NFTN)}NFTN*\ln\,(NFTN)}}$$

$$= \frac{e^{\beta_{NFTN}}e^{\ln\,(NFTN+1)^{\beta_{NFTN*\ln(NFTN)}(NFTN+1)}}}{e^{\ln\,(NFTN)^{\beta_{NFTN*\ln\,(NFTN)}NFTN}}}$$

$$= \frac{\left(e^{\beta_{NFTN}}\right)(NFTN+1)^{\left(\beta_{NFTN*\ln\,(NFTN)}\right)(NFTN+1)}}{NFTN^{\beta_{NFTN*\ln\,(NFTN)}NFTN}}.$$

No significant interactions were found. The AUC is 0.681, with a 78.3% hit rate for classification. The ROC curve and confusion matrix are shown in Figure 11 and Figure 12 respectively.

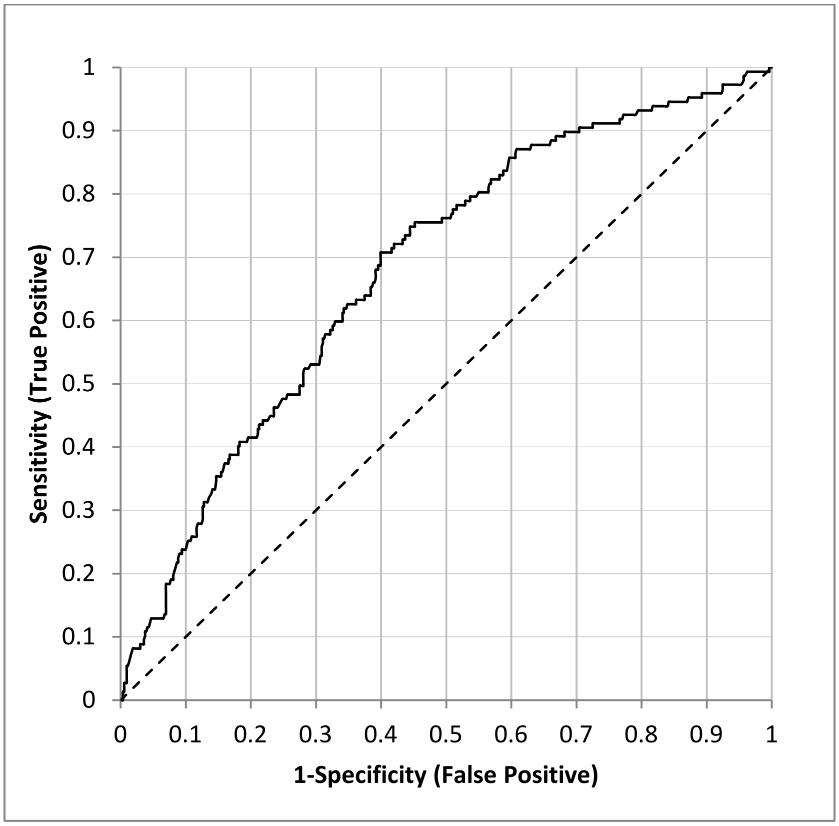


Figure 11. ROC Curve for Logistic Regression Damage Prediction Model. AUC = 0.681.

		Predicted	
		0	**1**
Actual	**0**	527	4
	1	143	4

Figure 12. Confusion Matrix for Logistic Regression Damage Prediction Model. Hit Rate = 78.3%.

Logistic Regression Human vs. Mechanical Error Model

The importance of human error as a failure cause was quantified by a model that recoded the dichotomous dependent variable FAILURE into a polytomous variable, FAILURE3, with three categories: 0 = Human Error-caused Failure, 1 = Mechanical-caused Failure, and 2 = No Failure. Mechanical-caused failures encompassed events where natural elements, autopilot errors, loss of communications, or electrical shorts led to SUAS failures. Human Error-caused Failures included pilot error, ground control operator error, or maintenance error which led to SUAS failures. Failures resulting from design errors were included in the Human Error category, despite the fact that they often produced effects that appeared to belong in the Mechanical-caused category.

A logistic regression model was constructed to classify each flight in the data set into one of the three categories. The model (n = 651, with prior probabilities Pr{FAILURE3 = 0} = 0.183, Pr{FAILURE3 = 1} = 0.184, and Pr{FAILURE3 = 2} = 0.633) has the parameter estimates shown in

Table **6**.

The ROC curves for the model (see Figure 13) have $AUC_0 = 0.700$, $AUC_1 = 0.750$, and $AUC_2 = 0.730$. The hit-rate on the confusion matrix (see Figure 14) is 64.8%. Nonlinearity was found in the logit, which was corrected with the addition of a Box-Tidwell term on (DSPLF + 1). The 1 was added to every instance of DSPLF since it often has values of 0, which would otherwise send its natural logarithm to negative infinity. A significant interaction was found between NFTOT and MAN, but the inclusion of this term lowered the classification accuracy, so it was not retained in the model.

Table 6. Parameter estimates and significance for Human vs. Mechanical Error Model.

	Term	Estimate	Std Error	Chi Square	Prob> ChiSq
Model for Human Error	Intercept	-4.68195	0.77784	16.49	0.000
	PROT	1.82054	0.32554	31.27	0.000
	NFTOT	0.00027	0.00058	0.22	0.638
	MAN	1.22563	0.42476	8.33	0.004
	MINWIND	0.08636	0.02980	8.4	0.004
	NFTN	0.05039	0.01390	13.13	0.000
	NFAF	-0.00602	0.00235	6.58	0.010
	DSPLF	0.08685	0.06932	1.57	0.210
	TEMP	0.01697	0.00870	3.81	0.051
	(DSPLF+1)*ln(DSPLF+1)	-0.02314	0.01829	1.6	0.206
Model for Mechanical Error	Intercept	-2.49574	0.76150	3.68	0.055
	PROT	1.83844	0.40826	20.28	0.000
	NFTOT	-0.00256	0.00063	16.82	0.000
	MAN	0.23041	0.53824	0.18	0.669
	MINWIND	-0.03502	0.03198	1.2	0.274
	NFTN	0.02523	0.01365	3.42	0.065
	NFAF	-0.00444	0.00204	4.72	0.030
	DSPLF	0.08163	0.03075	7.05	0.008
	TEMP	0.00917	0.00855	1.15	0.284
	(DSPLF+1)*ln(DSPLF+1)	-0.01366	0.00574	5.66	0.017

Table 7. Odds Ratios for the Human vs. Mechanical Error model

	Term	Odds Ratio	Lower 95%	Upper 95%
Model for Human Error	PROT	6.17522	3.26246	11.68842
	NFTOT	1.00027	0.99914	1.00140
	MAN	3.40630	1.48157	7.83154
	MINWIND	1.09019	1.02834	1.15576
	NFTN	1.05168	1.02341	1.08074
	NFAF	0.99400	0.98944	0.99858
	DSPLF	-	-	-
	TEMP	1.01712	0.99993	1.03460
	(DSPLF+1)*ln(DSPLF+1)	-	-	-
Model for Mechanical Error	PROT	6.28670	2.82427	13.99401
	NFTOT	0.99744	0.99622	0.99866
	MAN	1.25912	0.43844	3.61597
	MINWIND	0.96558	0.90691	1.02805
	NFTN	1.02555	0.99848	1.05334
	NFAF	0.99557	0.99160	0.99957
	DSPLF	-	-	-
	TEMP	1.00921	0.99243	1.02626
	(DSPLF+1)*ln(DSPLF+1)	-	-	-

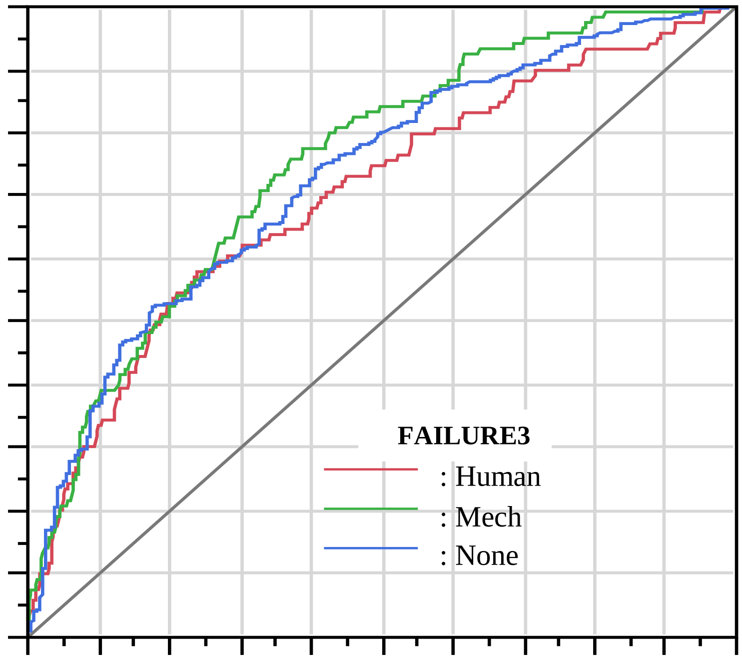


Figure 13. ROC Curve for Human vs. Mechanical Error Model.

		Predicted		
		0	**1**	**2**
Actual	**0**	16	9	94
	1	5	21	94
	2	8	19	385

Figure 14. Confusion Matrix for Logistic Regression Human vs. Mechanical Error Model. Hit Rate = 64.8%.

Artificial Neural Network Failure Prediction Model

For the ANN Failure Prediction Model, the response variable under consideration is FAILURE (0 = no failure, 1 = failure). The ANN Failure Prediction Model is designed primarily to screen out nonsalient features, and its input data (n = 539, with prior

probabilities Pr{FAILURE = 0} = 0.62 and Pr{FAILURE = 1} = 0.38) are a subset of the data used for the Logistic Regression Failure Prediction model. Since all variables are included in the baseline model and refinements are made to the model by sequentially removing variables, those variables with potentially problematic correlations were removed prior to architecture selection and model building. The correlation matrix for the input data was computed, resulting in the significant correlations shown in Table 8.

Table 8. Significant correlations for ANN input.

Factor 1	Factor 2	Correlation
NFTOT	NFAPT	0.833
DSTNLF	DSAFLF	0.702
MINWIND	MAXWIND	0.689

For interpretation reasons, NFAPT was removed from consideration. It is easier to track and interpret NFTOT than NFAPT. Likewise, MINWIND was removed from consideration. Its value to a decision maker is less than that of MAXWIND, as most regulations and safety requirements decree a maximum wind level at which an SUAS is allowed to operate. DSTNLF and DSAFLF are not expected to be significant in the model (based on the results of the logistic regression analysis) and are left in.

The input data was appended with a noise feature generated as a Uniform(0,1) random variate. The data was then randomized and partitioned into training, validation and test sets (70%, 15%, and 15%, respectively). Ninety feedforward ANNs with one hidden layer, 100 epochs maximum, and backpropagation training were constructed in MATLAB for each number of hidden layer nodes considered for the architecture. The

average test set misclassification rate was plotted as a function of the number of hidden nodes (see Figure 15) to determine a best network architecture. Based on the results of this analysis, a network with 18 hidden nodes was selected, as this provides the minimum number of nodes before the misclassification rate begins increasing.

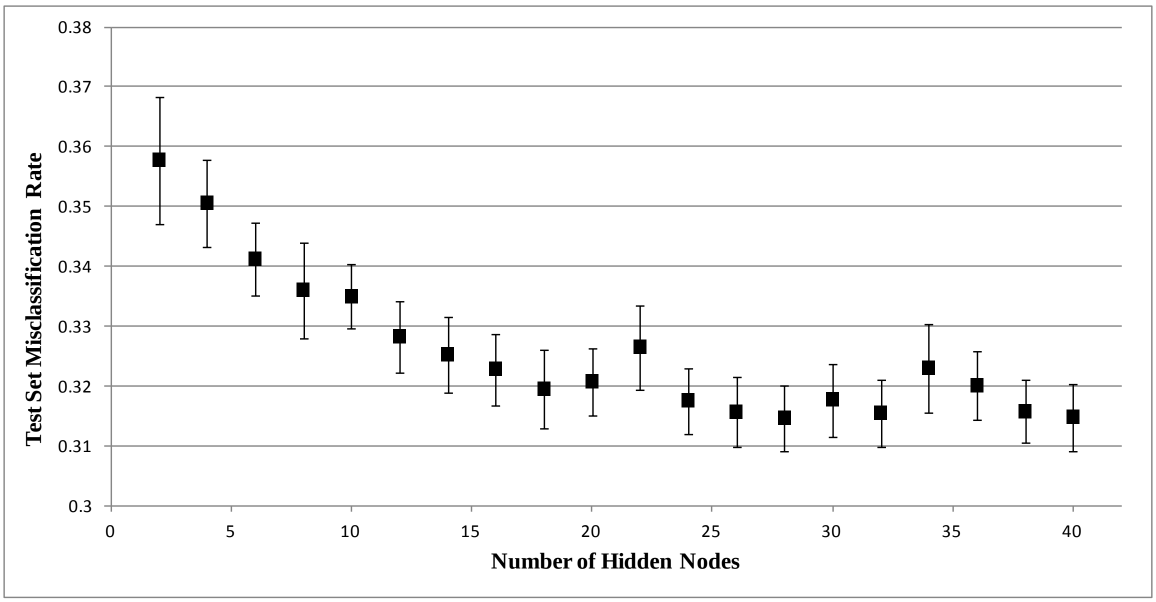


Figure 15. Test set misclassification rate as a function of number of hidden nodes for the ANN Failure Prediction Model (95% confidence interval).

Features were removed sequentially according to the SNR saliency criteria. Fifty feedforward neural networks each with one hidden layer and 18 nodes were trained to a maximum of 100 epochs, with backpropagation. The SNR saliency of each feature was computed after each network was trained and the least salient feature was denoted. The feature that received the most “least salient” rankings out of the 50 runs was removed. Features were removed in the order given in Table 9.

Table 9. Feature order of removal for the ANN Failure Prediction Model.

Total Number of Features Removed	Feature Selected for Removal
0	DSLOCLF
1	DSPLF
2	DSAFLF
3	DSAPTLF
4	DSTNLF
5	TIME
6	DSLF
7	DSLM
8	TEMP
9	MAXWIND
10	MAN
11	NFTN
12	NFP
13	NFAF
14	NFTOT
15	NFLOC
16	PROT
17	NOISE

The test set misclassification rate was plotted against the number of removed features for each of the 50 neural networks to determine the optimal number of features to retain in the model (see Figure 16). The plot shows the misclassification rate decreasing until 9 to 11 features are removed, after which the misclassification rate dramatically increases.

The three models, for 9, 10, and 11 features removed were compared against one another, with the noise variable removed from the input. One hundred neural network models were created using the same settings as before. The results are shown in Table 10. While a parsimonious model is desirable, so is an accurate model. The network that

removed 10 features appears to perform best, with the lowest minimum, lowest average, and only 7 features retained in the model. The variable MAN was retained in this model (but would not be retained in "model 11"), which made it attractive because MAN was declared earlier to be a potentially important control variable to include whenever possible.

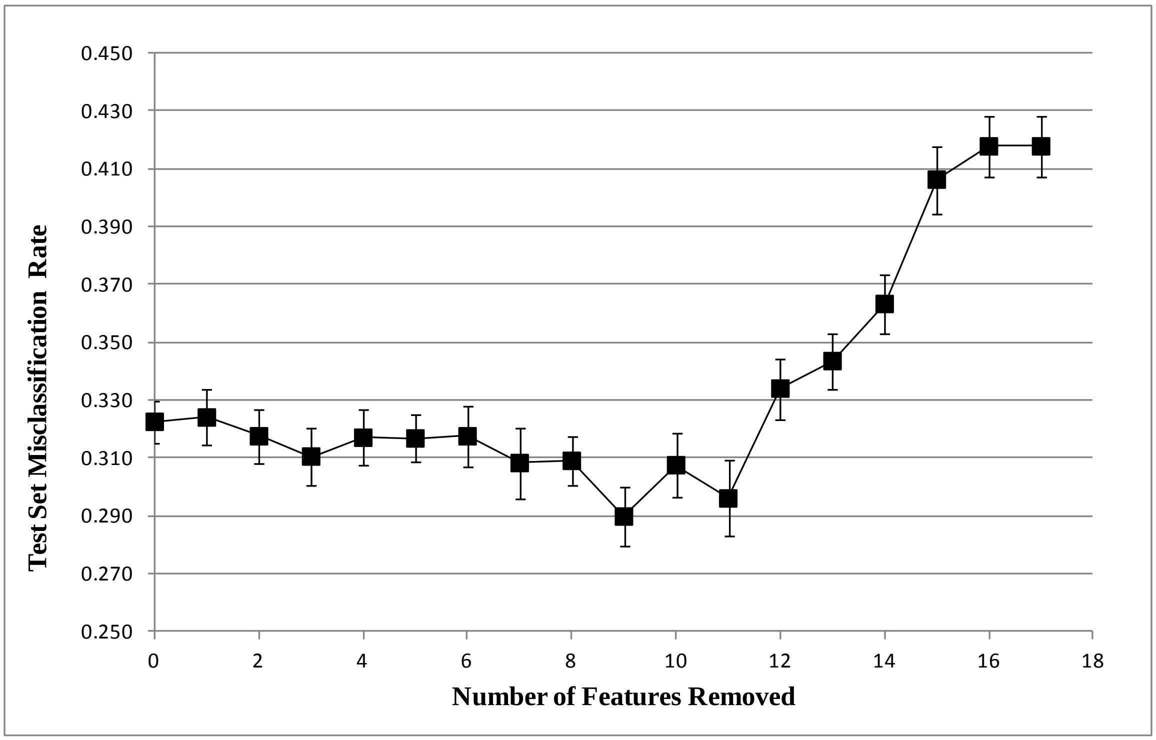


Figure 16. Test set misclassification rate as a function of features removed for the ANN Failure Prediction Model (95% confidence interval).

The architecture for the 7-feature model (containing features MAN, NFTN, NFP, NFAF, NFTOT, NFLOC, and PROT) was constructed and 100 networks using this structure were randomly initialized and simulated. From these resulting possibilities, the network model with the lowest observed test set misclassification rate was selected. This

model is the final ANN model selected for Failure Prediction. Its ROC curve and confusion matrix are shown in Figure 17 and Figure 18, respectively. The AUC is 0.724, with a 69.8% hit rate for classification.

Table 10. Comparison of three candidate ANN Failure Prediction models with 9, 10, and 11 features removed, results for 100 networks.

Features Removed	9	10	11
Average Test Set Misclassification Rate	0.300	0.297	0.305
Upper 95% Rate	0.307	0.304	0.312
Lower 95% Rate	0.293	0.291	0.299
Minimum Rate	0.238	0.225	0.225
Maximum Rate	0.500	0.375	0.388

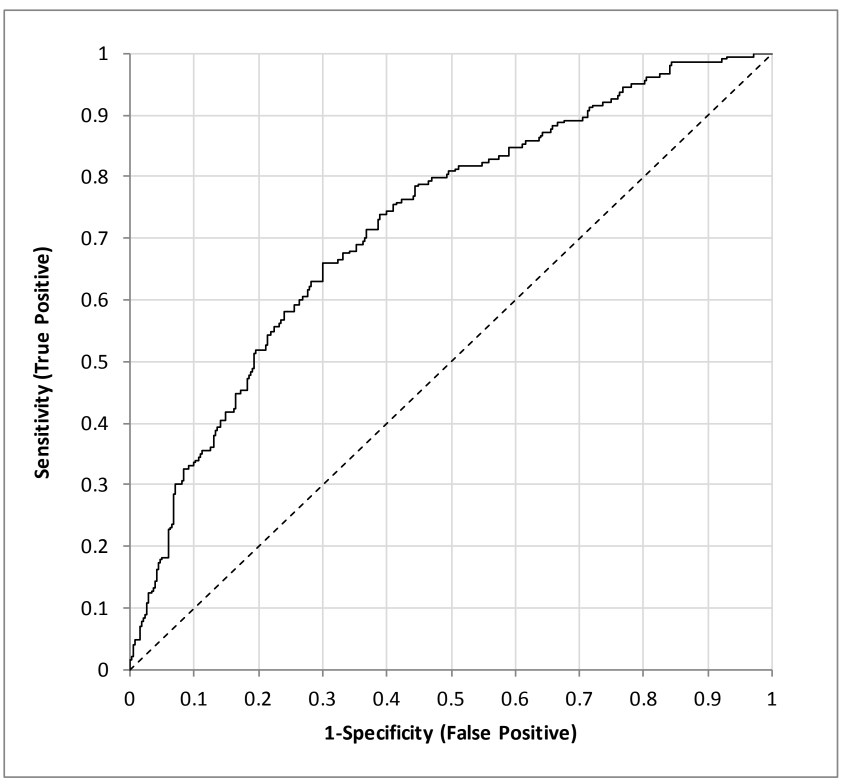


Figure 17. ROC Curve for ANN Failure Prediction Model. AUC = 0.724.

		Predicted	
		0	**1**
Actual	**0**	286	50
	1	113	90

Figure 18. Confusion Matrix for ANN Failure Prediction Model. Hit Rate = 69.8%.

Artificial Neural Network Damage Prediction Model

For the ANN Damage Prediction Model, the response variable under consideration is DAMAGE (0 = no SUAS damage, 1 = damage to SUAS). The input data (n = 539, with prior probabilities Pr{DAMAGE = 0} = 0.803 and Pr{DAMAGE = 1} = 0.197) was preprocessed the same way as it was for the ANN Failure Prediction Model. The architecture selection process was identically performed but was considerably more difficult as there was no clear "best" choice from the plot of test set misclassification rate versus number of hidden nodes (See Figure 19). The 20-hidden node structure was selected, as it appeared to have a low average misclassification rate and would maintain approximately the same structure as the previous model.

Features were removed sequentially according to the SNR saliency criteria. Fifty feedforward neural networks each with one hidden layer and 20 nodes were trained in MATLAB to a maximum of 100 epochs, with backpropagation. The SNR saliency of each feature was computed after each network was trained and the least salient feature was denoted. The feature that received the most "least salient" rankings out of the 50 runs was removed. Features were removed in the order given in Table 11.

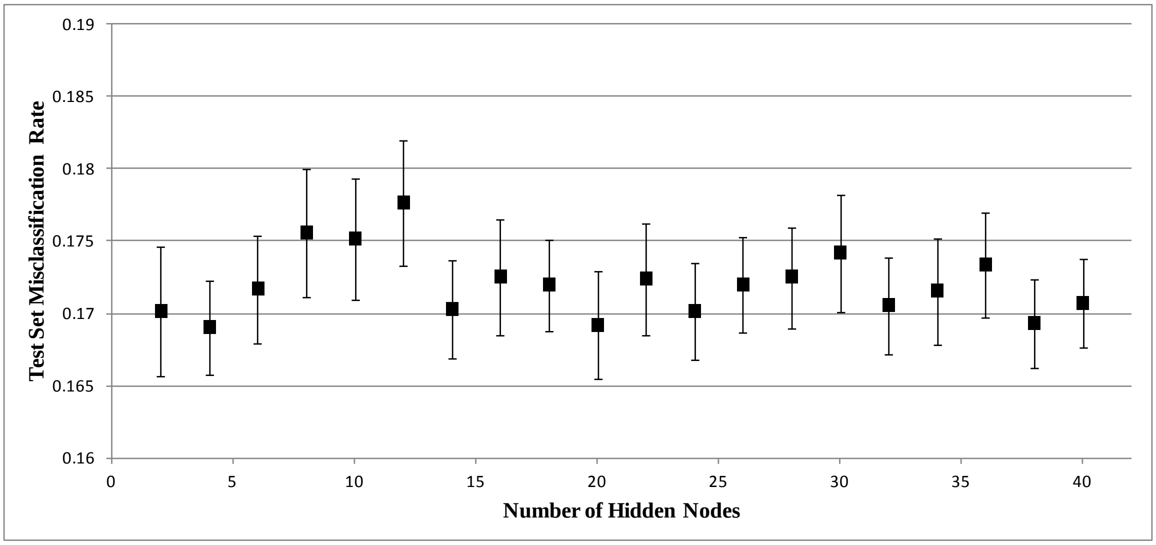


Figure 19. Test set misclassification rate as a function of number of hidden nodes for the ANN Damage Prediction Model (95% confidence interval).

The test set misclassification rate was plotted against the number of removed features for each of the 50 neural networks to determine the optimal number of features to retain in the model (see Figure 20). The plot shows the misclassification rate fluctuating until 11 to 13 features are removed, after which the misclassification rate dramatically increases, then decreases.

The three models, for 11, 12, and 13 features removed were compared against one another, with the noise variable removed from the input. One hundred neural network models were created using the same settings as before. The results are shown in Table 12. While a parsimonious model is desirable, so is an accurate model. The network that removed 12 features appears to perform best, with the lowest minimum rate, lowest average rate, and only 5 features retained in the model.

Table 11. Feature order of removal for the ANN Damage Prediction Model.

Total Number of Features Removed	Feature Selected for Removal
0	DSPLF
1	DSAFLF
2	DSAPTLF
3	DSTNLF
4	DSLOCLF
5	TIME
6	DSLM
7	DSLF
8	TEMP
9	MAN
10	NFTN
11	MAXWIND
12	NFP
13	NFTOT
14	NFAF
15	NFLOC
16	PROT
17	NOISE

Table 12. Comparison of three candidate ANN Damage Prediction models with 11, 12, and 13 features removed, results for 100 networks.

Features Removed	11	12	13
Average Test Set Misclassification Rate	0.163	0.160	0.162
Upper 95% Rate	0.165	0.162	0.163
Lower 95% Rate	0.162	0.159	0.161
Minimum Rate	0.138	0.125	0.150
Maximum Rate	0.175	0.175	0.188

The 5-feature model (containing features NFP, NFTOT, NFAF, NFLOC, and PROT) was simulated an additional 100 times at which point a model was selected with

the lowest observed test set misclassification rate. This model is the final ANN model selected for Damage Prediction. Its ROC curve and confusion matrix are shown in Figure 21 and Figure 22, respectively. The AUC is 0.742, with an 82.4% hit rate for classification.

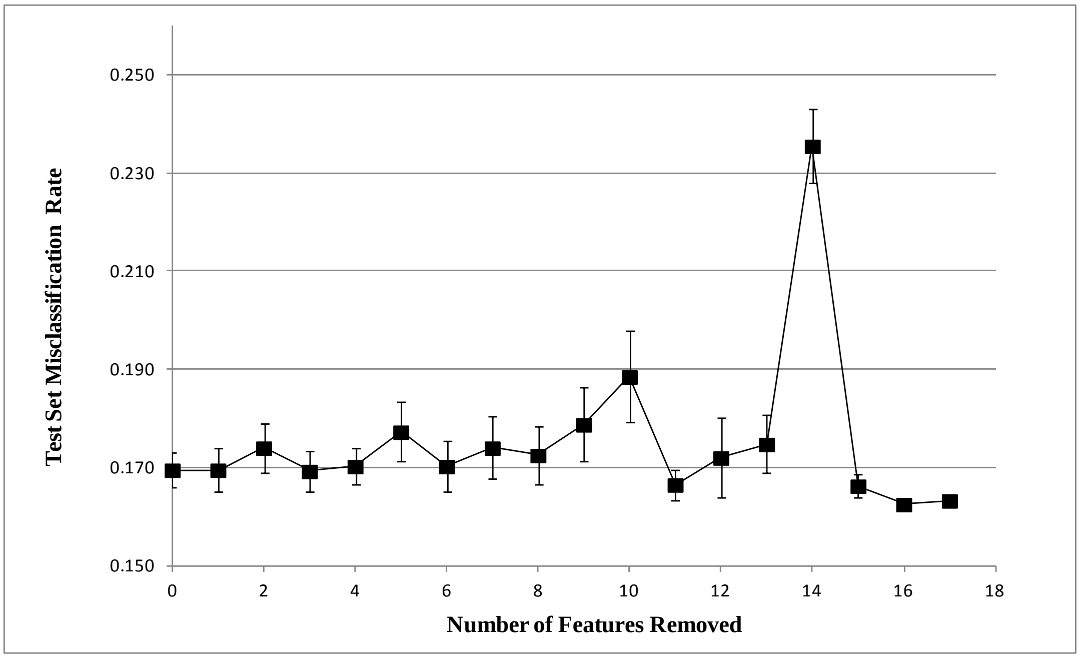


Figure 20. Test set misclassification rate as a function of features removed for the ANN Damage Prediction Model (95% confidence interval).

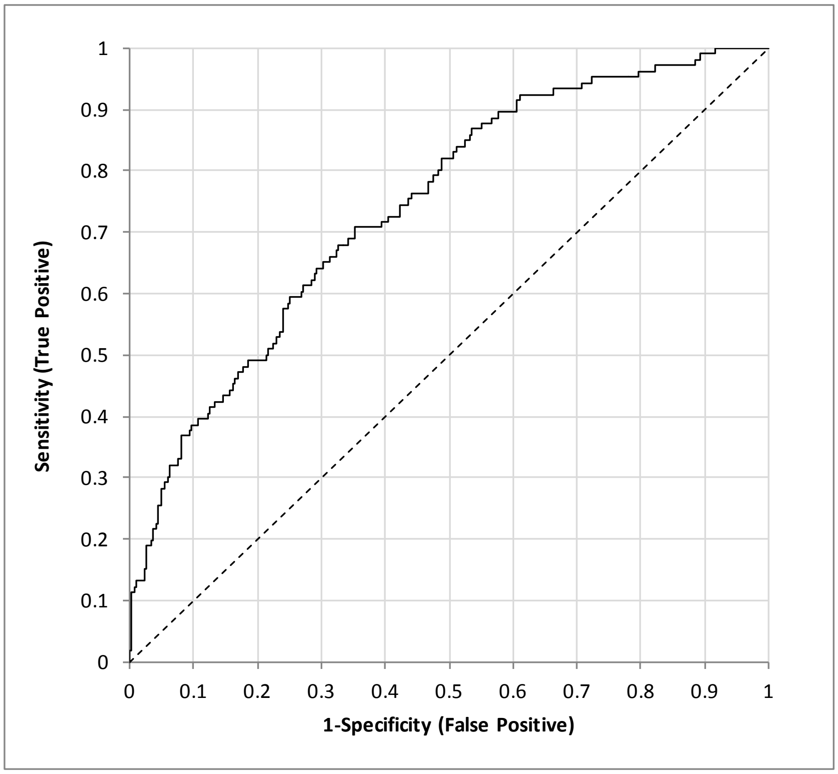


Figure 21. ROC Curve for ANN Damage Prediction Model. AUC = 0.742

		Predicted	
		0	**1**
Actual	**0**	425	8
	1	87	19

Figure 22. Confusion Matrix for ANN Damage Prediction Model. Hit Rate = 82.4%.

Artificial Neural Network Human vs. Mechanical Error Model

The same coding system from the Logistic Regression Failure Prediction Model was used to reclassify the dichotomous variable FAILURE into a polytomous variable, FAILURE3 with classes: 0 = Human Error-caused Failure, 1 = Mechanical-caused

Failure, and 2 = No Failure. Mechanical-caused failures included natural elements, autopilot errors, loss of communications, or electrical shorts. Human Error-caused Failures included pilot error, ground control operator error, design error or maintenance error.

A neural network model was constructed to classify each flight in the data set into one of the three categories. The model ($n = 539$, with prior probabilities Pr{FAILURE3 = 0} = 0.173, Pr{FAILURE3 = 1} = 0.204, and Pr{FAILURE3 = 2} = 0.623) was constructed in the same way (and with the same parameters) as the previous two neural network models. The architecture selection phase showed that 10 hidden layer nodes were optimal (see Figure 23). The feature screening phase suggested a closer look at the models with 10, 11 and 12 features screened (see Figure 24). The order of removed features is given in Table 13.

The three models, for 10, 11, and 12 features removed were compared against one another, after the noise variable had been removed from the input. One hundred neural network models were created using the same settings as before. The results are shown in Table 14. While a parsimonious model is desirable, so is an accurate model. The network that removed 11 features appears to perform best, with the lowest minimum rate, lowest average rate, and only 6 features retained in the model.

The 6-feature model (containing features NFTN, NFP, NFLOC, NFAF, NFTOT, and PROT) was simulated an additional 100 times at which point a model was selected with the lowest observed test set misclassification rate. This is the final ANN model selected for Human vs. Mechanical Error Prediction. Its ROC curves and confusion matrix are shown in Figure 25 and Figure 26, respectively. The ROC curves for the

model have $AUC_0 = 0.698$, $AUC_1 = 0.751$, and $AUC_2 = 0.769$, and the hit rate on the confusion matrix is 67.2%.

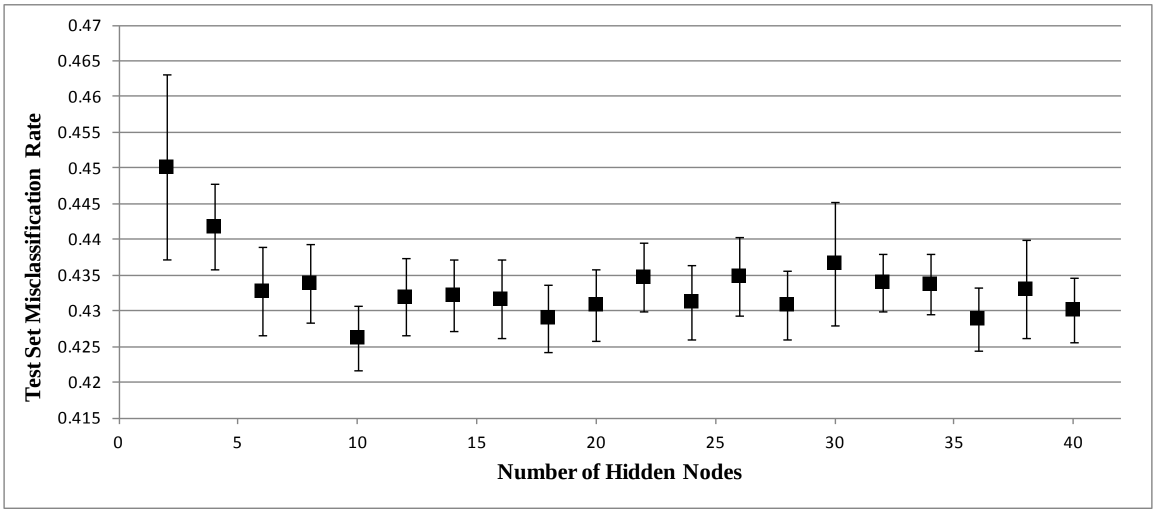


Figure 23. Test set misclassification rate as a function of number of hidden nodes for the ANN Human vs. Mechanical Error Model (95% confidence interval).

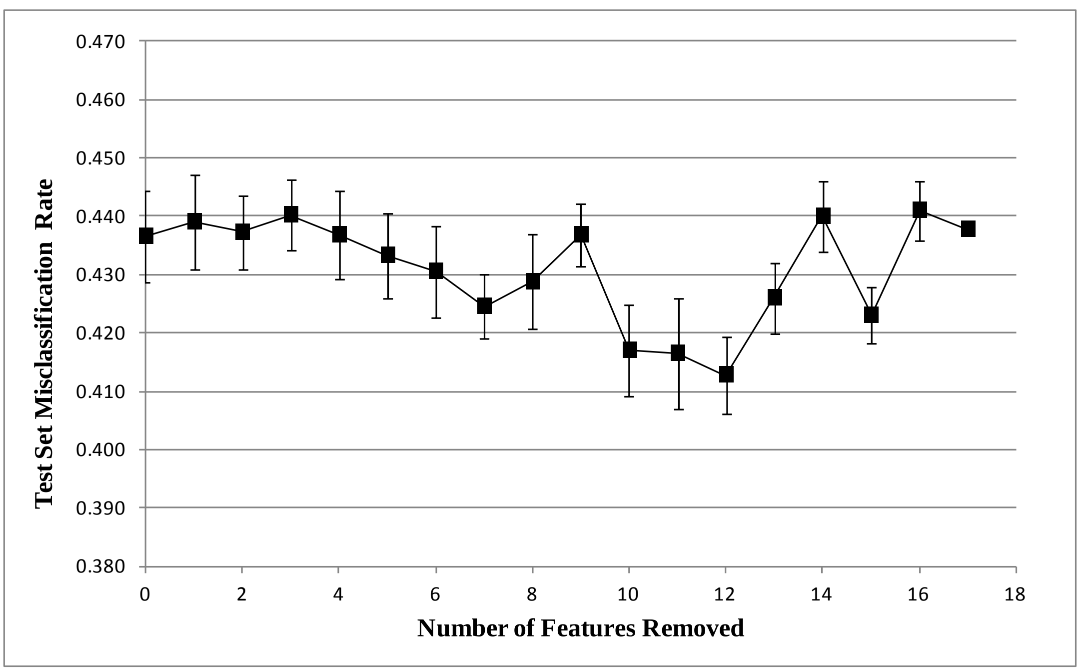


Figure 24. Test set misclassification rate as a function of features removed for the ANN Human vs. Mechanical Error Model (95% confidence interval).

Table 13. Feature order of removal for the ANN Human vs. Mechanical Error Model

Total Number of Features Removed	**Feature Selected for Removal**
0	DSLOCLF
1	DSAPTLF
2	DSAFLF
3	DSTNLF
4	DSPLF
5	MAN
6	TIME
7	DSLM
8	MAXWIND
9	DSLF
10	TEMP
11	NFTN
12	NFP
13	NFLOC
14	NFAF
15	NFTOT
16	PROT
17	NOISE

Table 14. Comparison of three candidate ANN Human vs. Mechanical Error Prediction models with 10, 11, and 12 features removed, results for 100 networks.

Features Removed	**10**	**11**	**12**
Average Test Set Misclassification Rate	0.426	0.413	0.422
Upper 95% Rate	0.432	0.419	0.428
Lower 95% Rate	0.419	0.408	0.416
Minimum Rate	0.363	0.350	0.375
Maximum Rate	0.588	0.525	0.538

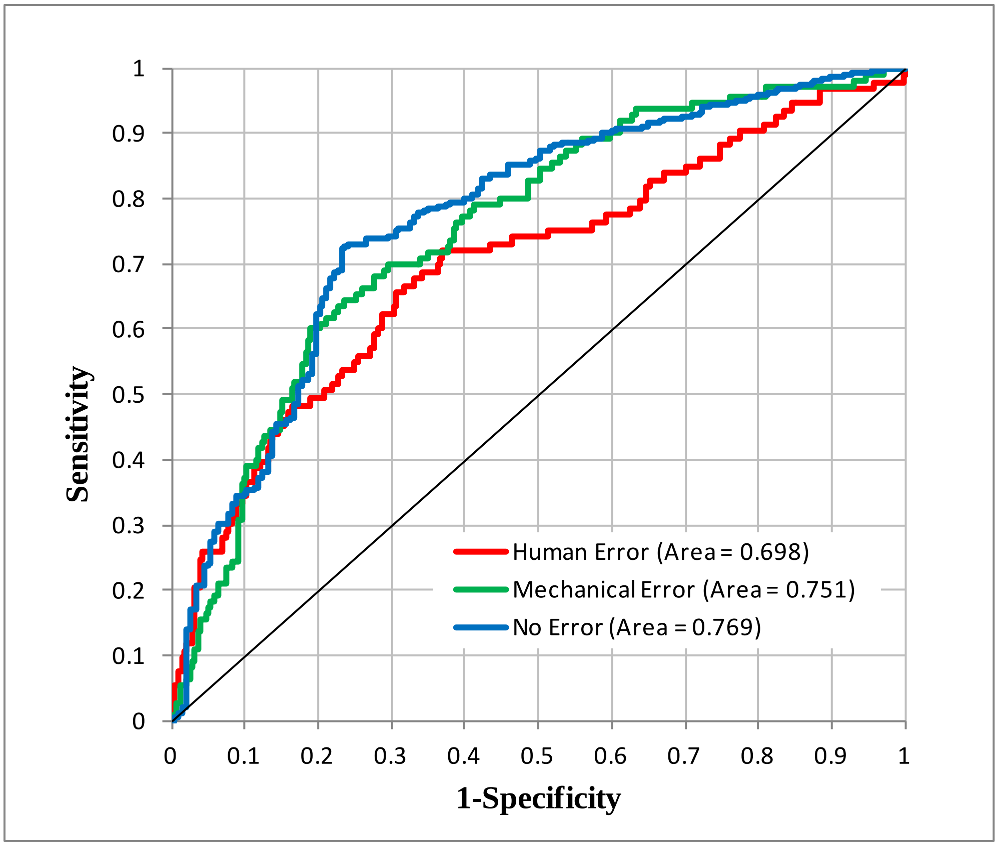


Figure 25. ROC Curve for Human vs. Mechanical Error Model.

		Predicted		
		0	**1**	**2**
Actual	**0**	22	13	58
	1	10	32	68
	2	8	20	308

Figure 26. Confusion Matrix for Human vs. Mechanical Error Model. Hit Rate = 67.2%.

IV. Results and Analysis

Logistic Regression Failure Prediction Model

The Failure Prediction Model is comparatively simple to analyze because it has neither interactions nor Box-Tidwell terms. The odds ratios are simply the exponentiation of the estimated parameters. The largest odd ratio is for the variable PROT, which indicates that, with all other model variables held constant, the choice to fly a prototype SUAS over a COTS SUAS increases the odds of a failure by 6 times. Likewise, with all model variables held constant, the same flight performed manually by a pilot has twice the odds of a failure as does that same flight with an autopilot.

Interestingly, while the other four variables had low odds ratios (near 1.0) all were significant at $\alpha = 0.05$. The one-unit increase may not the best metric for TEMP, because temperature is usually estimated at 5-degree intervals on AFRL/RWWV's flight reports. The five-unit increase odds ratio becomes $OR_{TEMP+5} = e^{5*0.01379} = 1.071$, which means that there is a 7% increase in the odds of a failure for every 5-degree temperature rise.

Similarly, the odds ratio of NFTOT is better computed for values larger than 1, since one flight out of the 854 total makes very little difference. In a similar computation as was done for TEMP above, the odds ratios for NFTOT are shown for multiple increments in Table 15. It shows that, with all other variables held constant, 10 flights of additional organizational experience decreases the odds of a failure by 1%. An additional 500 flights decreases those odds to about 60%.

Table 15. Odds ratio of NFTOT for multiple intervals

Additional Total Flights	Odds Ratio
5	0.995
10	0.990
25	0.975
50	0.951
100	0.904
250	0.777
500	0.603

In the failure prediction model, the variables NFAF and NFTN worked in opposition to one another. It was hypothesized that an increase in experience on a given aircraft would lead to greater operator and maintainer competency which would subsequently reduce failure rates. The model supports this hypothesis for airframes, but not for specific tail numbers. The opposite signs on NFAF and NFTN mean that more flights on an airframe (a given aircraft type, like the BATCAM or GENMAV) equate to lower odds of a failure, but more flights on a tail number (a specific vehicle like "BATCAM #12" or "GENMAV #3", for example) equate to higher odds of failure. This result is attributable to the fact that the vehicles are often flown to failure. While not all vehicles crash (and failures do not require damage to have occurred) a given tail number will fly until it is no longer needed for AFRL/RWWV's research or until it has crashed irreparably. In this dataset, a vehicle's last flight is usually a failure, so it is unsurprising that increases in NFTN positively correlate with failures. Additionally, there may be a physical basis for this result, as older vehicles may be more prone to mechanical failure. As with NFTOT above, increases in NFAF are given in more realistic intervals in Table 16. The results show that, all other values being equal, given the choice between two airframes, one should select the airframe with more flights to reduce the likelihood of

failure. An additional 10 flights lowers the odds of failure by about 5% and an additional 100 flights lowers the odds ratio to about 59%.

Table 16. Odds ratio of NFAF for multiple intervals

Additional Flights on Airframe	Odds Ratio
5	0.974
10	0.948
25	0.875
50	0.766
100	0.587

The overall performance of the failure prediction model is acceptable, with an AUC of 0.718 and a 69.6% hit rate for classification. Both measures indicate that the model outperforms simple guessing, but not by much. A guess of “no failure” on every flight would result in a hit rate of 64%, and is equivalent to the point in the upper right of the ROC curve. By selecting a desired sensitivity, the corresponding specificity can be obtained for the model. If AFRL/RWWV desired 80% sensitivity, the corresponding specificity is about 50%. For 90% sensitivity, the specificity drops to about 37%.

Some sample calculations may serve to better illustrate the operation of this model. Consider three hypothetical flights, whose data are shown in Table 17.

Table 17. Sample Calculation Data for Three Hypothetical Flights

Flight #	PROT	MAN	NFTOT	NFAF	NFTN	TEMP
1	0	0	250	50	15	72
2	1	0	250	50	15	72
3	1	0	500	50	15	72

Using the coefficient estimates from the model, we compute the logit for Flight #1as:

$$g_1(\vec{x}) = -2.67096 + 1.81659(0) + 0.74216(0) - 0.00101(250) - 0.00533(50) + 0.03781(15) + 0.01379(72)$$

$$g_1(\vec{x}) = -1.630.$$

The odds of a failure for Flight #1 becomes:

$$odds_1 = e^{g_1(\vec{x})} = e^{-1.630} = 0.1960.$$

Which makes the probability of a failure:

$$\pi_1 = \frac{e^{g_1(\vec{x})}}{1 + e^{g_1(\vec{x})}} = \frac{0.1960}{1 + 0.1960} = 0.164.$$

The model predicts a 16.4% probability of a failure given the Flight #1 values for the independent variables. If the same flight on the same day were flown by a prototype aircraft (PROT = 1, and assuming identical NFAF and NFTN) the data for Flight #2 would be used in the model. Following the same procedure shown above, the results would be:

$$g_2(\vec{x}) = 0.1867,$$

$$odds_2 = 1.205,$$

$$\pi_2 = 0.547.$$

The change from a COTS SUAS to a prototype SUAS increases the probability of a failure from 16.4% to 54.7%. If the model were left in its default state with a classification cutoff percentage of 50%, Flight #1 would be classified as a “No Failure” outcome and Flight #2 would be classified as a “Failure”. Note that the ratio of the odds for both flights is equivalent to the odds ratio for the variable PROT (the only variable that was altered) from Table 3,

$$OR_{PROT} = \frac{odds_2}{odds_1} = \frac{1.205}{0.1960} = 6.15.$$

Now consider Flight #3, which is identical to Flight #2 except that AFRL has now completed 500 total flights. Perhaps Flight #2 was canceled and the hypothetical prototype SUAS was placed on the shelf while 250 flights were accumulated, after which the same flight test was attempted. The results from calculations on Flight #3 are as follows:

$$g_3(\vec{x}) = -0.066,$$

$$odds_3 = 0.936,$$

$$\pi_3 = 0.484.$$

Flight #3 has a 48.4% probability of a failure, which would be classified as "No Failure". The odds ratio between Flight #3 and Flight #2 is:

$$OR_{NFTOT+250} = \frac{odds_3}{odds_2} = \frac{0.936}{1.205} = 0.777.$$

This is the same odds ratio that can be found in Table 15, which gave odds ratios for increases in NFTOT. Since the only difference between Flight #3 and Flight #2 was the 250 flight increase in NFTOT, the odds ratios between these two flights matches the value for 250 in the table.

Logistic Regression Damage Prediction Model

The model's significant terms and parameter estimates are comparable to those in the Failure Prediction Model except that TEMP and NFTOT were not found to be significant in the model when controlling for the other variables. The choice of a

prototype or COTS SUAS only affects the odds ratio by a factor of 4 rather than 6, and manual flight versus autopilot flight gives a multiple of 1.8 instead of 2. NFAF has the same relationship with damage as it did with failures: greater airframe experience led to lower odds on negative outcomes. See Table 18 for the odds ratio on NFAF at different intervals. The nonlinearity in NFTN meant that the odds ratio of NFTN varied, crossing above 1.0 at values over 10. Thus, NFTN behaved the same way as it did in the Failure Prediction Model for values greater than or equal to 10; more flights on a given tail number meant greater odds of a negative outcome. Interestingly, for less than 10 flights, the effect of each subsequent flight, up to flight number 10, was to decrease the risk of damage by lowering the odds ratio. Graphically, this is shown in Figure 27, where the odds ratio for one-unit increases in NFTN are plotted against NFTN's current value. An odds ratio of 1 is dashed in for reference.

The Damage Prediction Model performed poorly overall. Although the AUC was 0.681, the classification hit rate was only 78.3%. This is the same as the percentage of "no damage" outcomes in the dataset. Out of 678 flights, the model only predicted 8 "damage" flights, 4 of which were correctly classified. For 80% sensitivity, the model produces about 43% specificity. For 90% sensitivity, 30% specificity can be obtained.

Table 18. Odds ratio of NFAF for multiple intervals

Additional Flights on Airframe	Odds Ratio
5	0.959
10	0.919
25	0.810
50	0.656
100	0.431

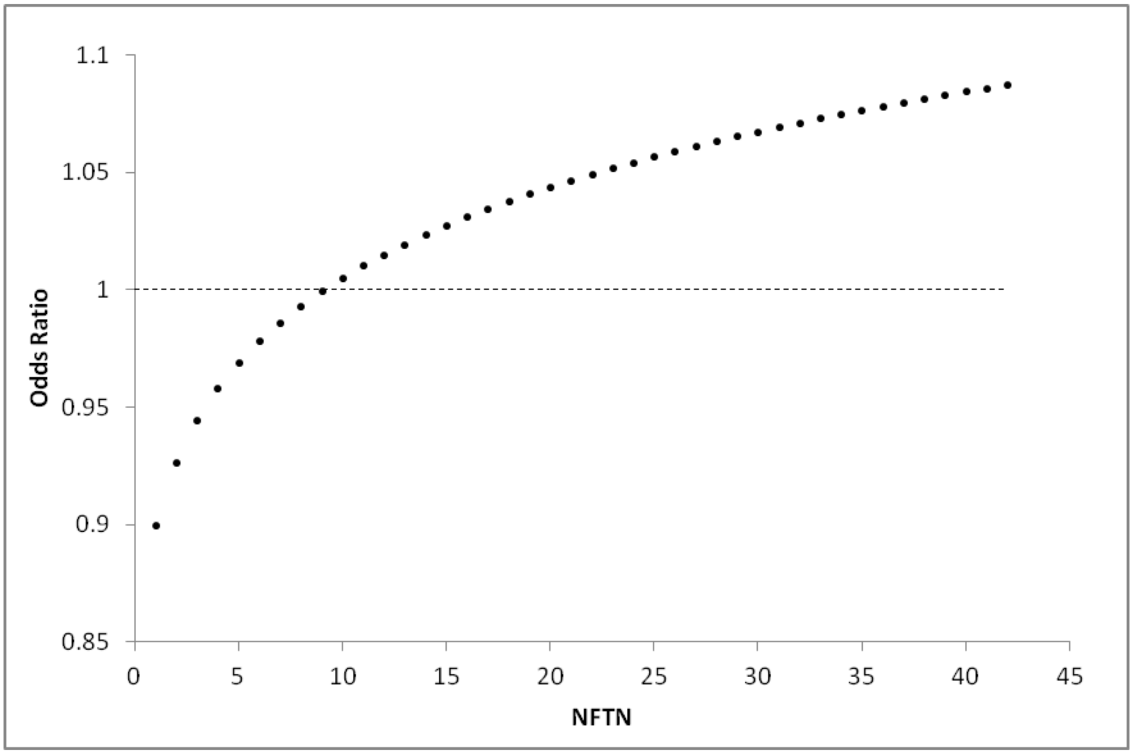


Figure 27. Odds Ratio for a one-unit increase in NFTN as a function of the present value of NFTN. Plotted across the range of NFTN values.

Logistic Regression Human vs. Mechanical Error Model

The Human vs. Mechanical Error Model is comparatively difficult to analyze as it not only contains a Box-Tidwell term, but it has a polytomous response variable with three levels. The whole model has two submodels, the first of which classifies between outcomes 0 (Human Error-caused Failure) and 2 (No Failure), and the second of which classifies between outcomes 1 (Mechanical Error-caused Failure) and 2 (No Failure). The classification function computes three probabilities, corresponding to outcomes 0, 1, and 2, the highest of which is selected as the estimated outcome.

To compute these probabilities, let $odds_0 = e^{g_0(\vec{x})}$ be the odds associated with the Human Error submodel for a given input vector, $\vec{x}$. Similarly, let $odds_1 = e^{g_1(\vec{x})}$ be the odds associated with the Mechanical Error submodel for the same input vector, $\vec{x}$. The probabilities of each outcome are computed as:

$$\pi_0 = \frac{odds_0}{1 + odds_0 + odds_1},$$

$$\pi_1 = \frac{odds_1}{1 + odds_0 + odds_1},$$

$$\pi_2 = \frac{1}{1 + odds_0 + odds_1}.$$

The Human vs. Mechanical Error model shares some similarities with the Failure Prediction Model. The parameter estimates for PROT compare well across both models, and indicate that the odds of a failure increase by a factor of between 6.2 and 6.3 when a prototype aircraft is selected over a COTS aircraft (holding all other variables constant). Because a nearly identical odds ratio affects both Human Error and Mechanical failure types nearly equally, this indicates that prototype aircraft are equally prone to mechanical as well as human error faults.

The choice of autonomous vs. manual flight (indicated by a 0 or 1, respectively in the MAN variable) was significant in the Human Error submodel (p-value = 0.0039) but was insignificant in the Mechanical Error submodel (p-value = 0.6686). Since the parameter estimate on MAN was 1.226 in the Human Error submodel, this meant that the odds of a Human Error-caused Failure increased by a factor of 3.41 when the SUAS was flown by a human rather than by the autopilot. Further, due to the insignificance of the MAN term in the Mechanical Error submodel, one cannot say with 95% confidence that

the choice between autonomous or manual flight affects the risk of a mechanical failure. This result accords well with theory and common sense.

The variable MINWIND was significant in the model, but only in the Human Error submodel. The model indicated than for every 1 knot increase in the minimum measured wind speed, the odds of a Human Error-caused failure increased by a ratio of 1.09. A five-knot increase would result in an odds ratio of 1.54, with all other variables held constant. Since MINWIND is not significant at $\alpha = 0.05$ on the Mechanical Error submodel, one cannot determine its effect on Mechanical-caused failures. This result is somewhat consistent with theory in that higher winds were expected to increase the risk of failures. It makes sense that higher winds could lead to more human error failures, especially in manual flight situations, but since environmental failures were lumped in with the mechanical category, it is at odds with theory that wind speed should be a poor predictor of mechanical failures as well.

The variables NFTN and NFAF worked as they did with the Failure Prediction Model. An increase in flights on a tail number is associated with an increased odds ratio of a failure. Meanwhile, an increase in flights on an airframe is associated with a decreased risk of failure. This was true in general for both submodels, (noting that NFTN was only significant to $\alpha = 0.065$ in the Mechanical Error submodel) and is a result of the fact that individual tail numbers are usually flown to failure and then eliminated from the flying population.

The variable DSPLF was significant in the model, and required a Box-Tidwell transformation term to linearize the logit. It was not significant in the Human Error submodel, but was very significant (both p-values < 0.02) to the Mechanical Error

submodel. This meant that while the pilot's currency (the number of days since the pilot had last flown) was important for classifying mishaps, it only impacted the classification when mechanical errors caused failures, and was not significant for human error failures. Since the model has two terms with DSPLF, it has a variable odds ratio dependent upon its current value, much like NFTN did in the Damage Prediction Model. The one-unit odds ratio has been computed for the Mechanical Error submodel, the only model for which DSPLF was significant and is shown in Figure 28. It shows that the odds ratio of a mechanical-caused failure is above 1 for low values of DSPLF, but decreases with successively larger values. This means that the more days a pilot has between flights (up to his 143rd day, which is the crossover with odds ratio = 1) the higher the risk of a mechanical-caused failure. Each missed day increases the risk of failure, but has less effect each successive day, until the 143rd day, after which each successive missed day lowers the risk of a mechanical failure.

The model is a poor classifier. While two of the three ROC curves are better than the Failure Prediction Model and all three are better than the Damage Prediction Model (measured by AUC), the overall classification accuracy from the Confusion Matrix shows a model just barely better than guessing. With 63.3% of flights ending in no mishap, the model was only able to correctly classify 64.8% of flights. The model only predicted 78 failures (when 239 had occurred) and, of those predicted, only 37 were correctly classified while 14 were classified as the wrong kind of failure. From the ROC curve, it can be seen that to achieve 80% sensitivity, only 42% specificity (for the lowest curve) is achieved. For 90% sensitivity, the model yields only 28% specificity.

Some sample calculations may serve to better illustrate the operation of this model. Consider three hypothetical flights, whose data are shown in Table 19. These are similar to the hypothetical flights from Table 17, except that the mean values of MINWIND and DSPLF are included in the independent variables, and the MAN variable is changed for the third flight rather than NFTOT.

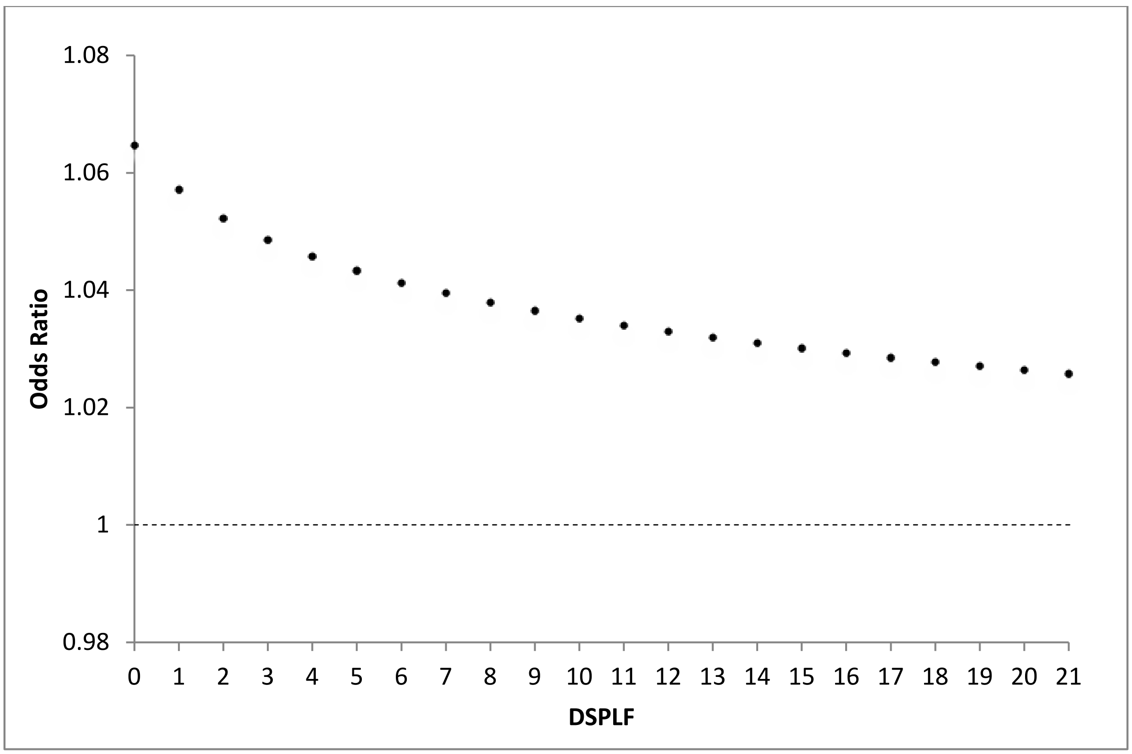


Figure 28. Odds Ratio for a one-unit increase in DSPLF as a function of the present value of DSPLF. Plotted for the 1 vs. 2 (Mechanical Error) Model for a three-week range.

Table 19. Sample Calculation Data for Three Hypothetical Flights

Flight #	PROT	NFTOT	MAN	MINWIND	NFTN	NFAF	DSPLF	TEMP
1	0	250	0	5	15	50	7	72
2	1	250	0	5	15	50	7	72
3	1	250	1	5	15	50	7	72

Using the coefficient estimates from the model, we compute the logit in the Human Error submodel for Flight #1 as:

$$g_{01}(\vec{x}) = -4.68195 + 1.82054(0) + 0.00027(250) + 1.22563(0) + 0.08636(5) + 0.05039(15) - 0.00602(50) + 0.08685(7) + 0.01697(72) - 0.02314(7+1)\ln(7+1)$$

$$g_{01}(\vec{x}) = -2.282.$$

The same is computed for the Mechanical Error submodel, yielding:

$$g_{11}(\vec{x}) = -2.151.$$

The odds of a Human Error-caused failure for Flight #1 becomes:

$$odds_{01} = e^{g_{01}(\vec{x})} = e^{-2.282} = 0.102.$$

The odds of a Mechanical Error-caused failure for Flight #1 becomes:

$$odds_{11} = e^{g_{11}(\vec{x})} = e^{-2.151} = 0.116.$$

Which makes the probability of a Human Error-caused failure:

$$\pi_{01} = \frac{e^{g_{01}(\vec{x})}}{1 + e^{g_{01}(\vec{x})} + e^{g_{11}(\vec{x})}} = \frac{0.102}{1 + 0.102 + 0.116} = 0.084.$$

For Mechanical Error-caused failures, the probability is:

$$\pi_{11} = \frac{e^{g_{11}(\vec{x})}}{1 + e^{g_{01}(\vec{x})} + e^{g_{11}(\vec{x})}} = \frac{0.116}{1 + 0.102 + 0.116} = 0.095.$$

For the "No Failure" outcome, the probability is:

$$\pi_{21} = \frac{1}{1 + e^{g_{01}(\vec{x})} + e^{g_{11}(\vec{x})}} = \frac{1}{1 + 0.102 + 0.116} = 0.821.$$

The model selects the outcome with the highest probability ($\pi_{21} = 82.1\%$), and would classify this as a "No Failure" flight. If the same flight were flown by a prototype

SUAS instead of a COTS SUAS (as in Flight #2 from Table 19, assuming the same pilot and identical NFTN and NFAF on the aircraft), the results change as follows:

$$g_{02}(\vec{x}) = -0.462,$$

$$g_{12}(\vec{x}) = -0.313,$$

$$odds_{02} = 0.630,$$

$$odds_{12} = 0.731,$$

$$\pi_{02} = \frac{0.630}{1 + 0.630 + 0.731} = 0.267,$$

$$\pi_{12} = \frac{0.731}{1 + 0.630 + 0.731} = 0.310,$$

$$\pi_{22} = \frac{1}{1 + 0.630 + 0.731} = 0.423.$$

The classification remains the same, "No Failure", with $\pi_{21} = 42.3\%$ as the highest of the three probabilities. If the same flight test with the same prototype aircraft was performed with a pilot flying manually rather than by autopilot (Flight #3 from Table 19, where MAN = 1) the model calculations produce the following results:

$$g_{03}(\vec{x}) = 0.764,$$

$$g_{13}(\vec{x}) = -0.082,$$

$$odds_{03} = 2.147,$$

$$odds_{13} = 0.921,$$

$$\pi_{03} = \frac{2.147}{1 + 2.147 + 0.921} = 0.528,$$

$$\pi_{13} = \frac{0.921}{1 + 2.147 + 0.921} = 0.226,$$

$$\pi_{23} = \frac{1}{1 + 2.147 + 0.921} = 0.246.$$

The classification category changes for Flight #3. Rather than "No Failure" like the previous two flights, the model selects "Human Error" as the most likely outcome for the flight. This accords with the interpretation of the coefficients and odds ratios provided above which indicated that prototype aircraft increase the risk of both types of failures, and that manual flight increases the risk of Human Error-caused failures. In the example case, the change to manual flight had a significant impact on the odds ratio to indicate a Human Error outcome. The change to manual flight will always have that impact on the odds ratios, but those odds ratios affect the classification outcomes differently, depending on the starting probabilities.

Artificial Neural Network Models

The ANN models were used to screen features to determine which factors had the greatest impact on each of the SUAS outcomes. The three ANN models showed many similarities with each other and had much in common with the logistic regression models, but the differences between them can also be exploited to gain insight into SUAS failures and damage.

The worst-performing ANN model was the model for Damage Prediction, which matches the results for logistic regression. The way the neural network model was formed gives additional insight into why both Damage models are barely better than guessing. When exploring a neural network's potential architecture, one should see a decreasing misclassification rate as more nodes are added until the rate stabilizes, and any additional nodes fail to provide improved performance. This behavior is clearly seen in the ANN

Failure Prediction Model (see Figure 15). The ANN Damage Model (see Figure 19) does not display this behavior. This means that there is no optimal, minimal architecture, which is probably a result of having data that is indistinguishable from noise. This can be seen in the plot of misclassification rate as a function of features removed, which should look like a mirror image of the architecture exploration plot. The ANN Damage Prediction Model (see Figure 20) has a fluctuating mishap classification rate, which spikes when most features are removed and then rapidly decreases. Since the last data point shows the misclassification rate when a noise variable is the only input to the model, it suggests that the other data add little to the classification, because including them results in a higher misclassification rate. Thus, both Damage Prediction Models are difficult to correctly classify based on the noise-like quality of their input data, relative to the output.

The ANN Human vs. Mechanical Error Model is better by comparison, exhibiting misclassification rate curves that are more typical of well-classifying neural networks. The model performed almost identically to the Logistic Regression Model, with ROC AUCs that nearly matched for both Human and Mechanical Error. This was an encouraging result and suggested that prediction of the specific types of error is possible. The confusion matrix clarifies that when Mechanical Error is predicted, the model only classifies it correctly $\frac{32}{13+20+32} = 49.2\%$ of the time. Interestingly, if the Human and Mechanical Error classifications are lumped together, this model has a higher hit rate than the Failure Prediction model, correctly predicting $\frac{22+13+10+32+308}{539} = 71.4\%$ of total

mishaps, regardless of cause. This outperforms the ANN Failure Prediction model, which only had a hit rate of 69.8%.

The ANN Failure Prediction Model was the best-looking model from an architecture selection and feature-screening perspective. It displayed the expected characteristics of a good classifying network. Its performance was very similar to that of the Logistic Regression Failure Prediction Model, with a nearly identical ROC curve and confusion matrix (when taking into account the different sample sizes). The features selected for the model compare favorably with those found by logistic regression, and nearly identically match those selected for the other ANN models. Table 20 presents the selected features for each model, ranked by order of significance (using p-value for logistic regression, and reverse order of removal for ANN).

Table 20. Feature ranking for all models. (*Asterisk denotes a transformed feature)

Failure Prediction		**Damage Prediction**		**Failure: Human vs. Mech Error**	
Log. Reg.	**ANN**	**Log. Reg.**	**ANN**	**Log. Reg.**	**ANN**
PROT	PROT	PROT	PROT	PROT	PROT
NFTN	NFLOC	NFAF	NFLOC	NFTOT	NFTOT
NFAF	NFTOT	NFTN*	NFAF	NFTN	NFAF
NFTOT	NFAF	MAN	NFTOT	MINWIND	NFLOC
TEMP	NFP		NFP	DSPLF*	NFP
MAN	NFTN			NFAF	NFTN
	MAN			MAN	
				TEMP	

Model Comparison

There are some consistencies in the features selected by all models. Most important, the variable PROT was the most significant to each model, for predicting both failure and damage. Clearly the single greatest indicator of flight outcome is whether the SUAS flown is a prototype model constructed by AFRL or a COTS model purchased from a manufacturer.

NFAF is the only other variable to appear in every model. This means that AFRL's experience with a given airframe is important to predicting flight outcomes. Likewise, NFTN and NFTOT appear in five of the six models, which suggests that they are significant factors to investigate for failure prevention. MAN was the next most important factor, appearing in four models, and is likewise worth noting for further analysis and investigation.

The preponderance of factors that begin with "NF" (and the corresponding dearth of terms beginning with "DS") indicates the importance of experience over intervals in determining SUAS flight outcomes. The "NF" factors record the total number of flights for each measure, which is a good approximation for overall experience (NFTOT for organizational experience, NFP for individual pilot experience and so forth). The "DS" factors record the days since an event occurred, which marks the intervals between events. These "DS" terms, with one exception, are surprisingly absent from this ranking.

The poor performance of the Damage Prediction Model (for both Logistic Regression and Artificial Neural Networks) casts suspicion on the important factors it suggests. If those two damage models are excluded from consideration, the remaining

models for Failure Prediction suggest PROT, NFAF, NFTN, NFTOT, and MAN as the most significant factors to address. Interestingly, some factors were favored by the different model-building tools, with Logistic Regression using TEMP exclusively and Artificial Neural Networks using NFLOC exclusively in both Failure models. These variables may also warrant consideration, but are less likely to be of practical importance, both from a physical perspective and from a modeling perspective.

Model Validation

The best-performing model, the Logistic Regression Failure Prediction Model, is investigated for validity using 50 flights from the first quarter of calendar year 2010. These data were not used in the construction of the model. Of the 50 flights, only 41 have complete input data and failure outcomes. There were 5 flights terminating in failures over this time period (12.2% failure rate), with three occurring on the same day, while trying to accomplish the same highly complex, high-risk (in the opinion of the test engineer) flight objective.

The model predicts 0 individual flight failures over the same period. See the Confusion Matrix in Figure 29. NFTOT has a large influence at approximately 900 flights, producing an odds ratio of $\boldsymbol{e^{-.00101*900} = 0.402}$. The most likely type of flights that can cause the model to predict failures at this high level of NFTOT are those with prototype SUAS being flown manually. No flights with these characteristics were attempted during this period. In order for the model to predict a failure for a COTS aircraft, the SUAS has to be flown manually and must have an NFTN in excess of 57.

This is unrealistic as Table 1 shows that the highest NFTN for the main dataset is 42. Thus, the model will probably not predict failures for COTS aircraft with a large NFTOT.

		Predicted	
		0	**1**
Actual	**0**	36	0
	1	5	0

Figure 29. Confusion Matrix for Validation of Logistic Regression Failure Prediction Model. Hit Rate = 87.8%.

Over this period, 83.0% of flights were COTS SUAS, compared to the 28.8% historical average and 57.2% for the same quarter of the previous calendar year. This indicates that the validation dataset does not reflect the typical composition of the historical data, but indicates a trend away from testing prototype aircraft.

Of interesting note, the two failures not associated with the high-risk flight test both occurred to COTS SUAS while under manual control. In both cases, the failure prediction probability was elevated due to flying under manual control. In one case, the particular SUAS had a large number of prior flights (NFTN = 38), which additionally raised its predicted probability of a failure. This reinforces the validity of MAN as a critical factor to be addressed for failure prevention, and suggests that NFTN may likewise be important.

Further, the model predicts the probability of failure for each of the 41 flights. Over the flights examined, the minimum probability of failure is 5.1%, the maximum is 41.5% and the median and mean are 10.2% and 13.7%, respectively. The model does not

predict that any specific flights will be failures (no individual probability is greater than 50%). Since there are 41 flights where there are non-zero (and sometimes significant) probabilities of failure, this suggests that the test engineer can expect a certain number of flight failures.

Assume that a decision maker or test engineer can specify these 41 flights in advance and wants to know the expected number of mishaps, given all the probabilities across all flights. The Poisson-Binomial distribution is examined, which gives the expected probability for a given number of failures occurring out of the 41 flights. In general, the Poisson-Binomial is the convolution of n independent, non-identical Bernoulli trials (Wang 1993). Each flight represents a non-identical Bernoulli trial, because it is a single trial with a unique probability of failure (assessed by the failure prediction model). The outcome "failure" is substituted where the word "success" would normally appear in the description of a Bernoulli trial, because "failure" is the outcome that is positively predicted by the model. The Poisson-Binomial can be solved iteratively using equations from Chen, Dempster and Liu (1994):

$$R(k, C) = \frac{1}{k} \Sigma_{i=1}^{k} (-1)^{(i+1)} T(i, C) R(k - i, C)$$

where

$R(k, C)$ is the probability of obtaining k "failure" trials,

$R(k - i, C)$ is the probability (previously computed) of obtaining $k - i$ "failure" trials,

and $T(i, C)$ is computed as shown:

$$T(i, C) = \Sigma_{j=1}^{n} w_j^i.$$

$R(0, C) = \Pi_{j=1}^{n} (1 - \pi_j)$ is the probability associated with zero "failure" trails,

$w_j = \left(\frac{\pi_j}{1-\pi_j}\right)$ is the odds of a failure on flight j,

and π_j is the probability of "failure" on each flight j out of n total flights.

This iterative equation was implemented in MATLAB to compute the expected number of failures for the 41 validation flights. The Poisson-Binomial distribution for these flights (see Figure 30) shows that five failures is the largest of all the binomial probabilities at 18.5%. Six failures and four failures are the next most likely, with 17.7% and 15.5% probabilities, respectively.

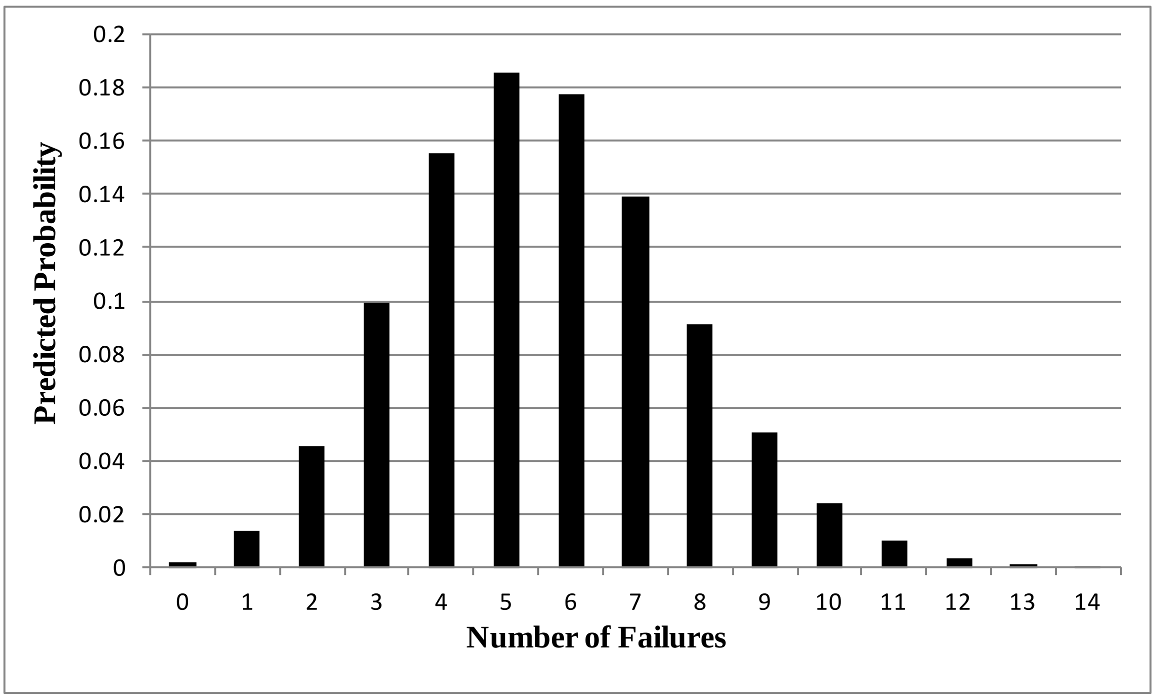


Figure 30. Poisson-binomial distribution for total number of failures given 41 flights with individual flight probabilities determined by logistic regression failure model.

The logistic regression failure prediction model, while not predicting any individual flight failures, nevertheless predicted (via the Poisson-Binomial distribution)

that five failures was the most likely outcome of the 41 flights, a result that is exactly validated by the dataset, in which five failures occurred. There is a wide range of statistical validity with such a small validation set, though. The bounds of a two-tailed 90% confidence level include outcomes from three to nine failures, and the bounds of a two-tailed 97% confidence level include outcomes from two to ten failures. Assuming a 97% confidence level, if the 41 flights result in two to ten total failures (inclusive), it will not be rejected as statistically different from the Poisson-Binomial model. Since there were five failures observed over these 41 flights, it can be concluded from these results that there is not enough statistical evidence to reject the validity of this model for predicting the expected total number of failures.

Model for Flight Planning

Using the results of the logistic regression modeling, a basic flight planning model can be constructed that provides decision makers with an estimated minimum number of flights to meet a given probability of success. The test engineers outline their objectives, select the SUAS platform to complete it, select a pilot to fly the mission, and collect all the necessary data as input for the logistic regression failure model. The output from this model provides an estimate of the probability of a failure for the given set of inputs. Assuming that failures result in complete data loss, this probability can be used to compute the minimum expected number of flights necessary to reach a given probability of overall mission success.

Since the same platform, test site, flight crew, and other associated variables will be used to achieve a specific flight objective on a given test day, the probability of an SUAS failure is assumed to be constant for that mission on that day. This is not entirely accurate, as each additional flight adds to NFTOT and NFAF (and NFTN if the same tail number is recycled). But these variables affect the odds ratio so slightly (and NFAF and NFTN work against one another) that the effect from flight to fight is small. In practice, the probabilities of failure of sequential flights with the same SUAS typically vary by less than 0.3% from flight-to-flight. Therefore, the output from the logistic regression failure model is a good approximation for the probability of failure across all flights on a given test day.

Let the failure probability, π, be the output of the logistic regression failure model and let the minimum probability of mission success, p, be determined by the decision maker or test authority. The minimum necessary number of flights flown, n, that are expected to meet this minimum success level is related to these probabilities as shown:

$$1 - \pi^n \geq p.$$

Given a minimum required probability of success and a probability of failure from the logistic regression model, the minimum expected number of flights can be computed as shown:

$$n \geq \frac{\ln(1 - p)}{\ln(\pi)}.$$

This is equivalent to computing the number of trials necessary for the sum of all binomial probabilities greater than zero to exceed probability p given π.

For example, assume that Flight #3 from Table 17 is specified by the test engineers. The logistic regression failure model predicts a probability of failure of 0.484 or 48.4%. This flight would be classified as a “No Failure” flight, but it still has a fairly high chance of failure. If a minimum probability of success of 90% was desired, the minimum expected number of flights is:

$$n \geq \frac{\ln(1 - 0.90)}{\ln(0.484)},$$

$$n \geq 3.17,$$

$$n = 4.$$

This means that when the probability of each flight failing is 48.4%, the test engineer can expect that all mission objectives will be achieved (all necessary flight data collected) 90% of the time if at least four flights are attempted. If the minimum probability of success is raised to 95%, the minimum number of flights is five. Obviously, a 100% success rate is theoretically impossible.

This model, while simplistic, provides a good rule of thumb for the test engineer to estimate the number of flights necessary to gather all the data. This model does suffer from a few shortcomings, though. As discussed, it assumes that the probabilities for each flight are constant, whereas they will vary slightly with the changes in NFTOT, NFAF, and NFTN on each successive flight. Further, a flight failure does not necessarily mean that all data is lost. If the failure occurs immediately upon takeoff, it is likely that all data for the test will be lost. If the failure occurs midway through, it is possible that some data could be salvaged, without having to be repeated by subsequent tests. To account for this, the test engineer may find that a minimum probability of success set closer to 80% or

lower works best, due to the partial gathering of data on each failure flight. Likewise, the complexity of the mission objectives may affect the estimated number of flights: a mission to see if a new launch capability performs correctly is a simple test whose result is known if the SUAS takes off, whereas a series of climbs and glides to assess engine and aerodynamic performance is more complex. The former may require a low probability of success in order for the model to reflect empirical results, whereas the latter may require a much higher probability. The occurrence of damage and its impact on this flight planning model is not addressed, but would also affect the number of flights, by possibly altering which aircraft could fly. If a damaged aircraft is replaced, the probability of a failure from the logistic regression failure model could change dramatically due to differences in NFAF (if a different model was selected for the mission) and NFTN (if a different tail number of the same model was selected).

V. Discussion

Summary

This research sought to determine if SUAS flight test failures and airframe damage could be predicted from parameters measured prior to flight. A failure was defined as a flight test terminating unexpectedly prior to all data being collected, regardless of the cause of the termination. Damage was defined as any injury to the airframe, regardless of cost or repair time. Both failures and damage were modeled with logistic regression to determine the quantifiable effects of each important parameter, and with artificial neural networks to provide an alternative method of parameter screening.

A review of the literature on large SUAS and manned aircraft mishaps (which are comparable to a composite of SUAS “failures” and “damage”) suggested that human error would be a leading cause of SUAS failures, and that increased pilot experience and currency would help reduce those failure rates. In the course of analysis, human error was found to be equally as prevalent as mechanical error, while pilot experience and currency were not found to significantly affect failure rates. Likewise, surface wind speed was hypothesized to affect failure rates, but this parameter, too, was not found to significantly affect observed failure rates. The one area where large UAS and manned aircraft results overlapped with the SUAS results obtained in this research is in the effect of experience. Mishap rates tend to decrease in the manned and large UAS communities over time as more flight hours are built up and as organizations adapt. So, too, did SUAS failure rates

decrease, both with total number of flights across all platforms, as well as with total flights for each type of airframe.

The overall results of the logistic regression analysis were that damage could not be accurately predicted, but failures could be. The neural network analysis confirmed that the measured parameters modeled damage no better than random noise. For failure modeling, the five main parameters deemed important by the logistic regression and the artificial neural network modeling merited further investigation for failure prevention.

The models developed from this data were not all equally useful. Most noticeably, the damage prediction models performed poorly as classifiers. This means that SUAS damage is not possible to predict with any greater accuracy than simple guessing, given the measured variables that were available. Damage appears to be a random outcome, with no discernible root causes. The primary conclusion regarding damage is that it occurs in about 23% of flights, with no clear preventive measures available. Discriminating between human-caused and mechanical-caused failures shows some promise, but the significant factors identified by the two modeling approaches were dissimilar and the prediction hit rates were weak.

The simple outcomes of “failure” and “no failure”, on the other hand, tend to be more predictable. There are common features that are correlated with the occurrence of SUAS failures that can be exploited to minimize future mishap rates. Two dichotomous variables, PROT (which indicated whether an SUAS was a lab-developed prototype or a Commercial-off-the-Shelf aircraft) and MAN (which indicated whether an SUAS was flown manually or with autopilot control) were significant in predicting failures. From the analysis, it can be concluded that, controlling for all other significant factors, flying a

prototype SUAS increases the odds of a failure by a factor of six over a COTS aircraft. Additionally, controlling for all other significant factors, flying an SUAS manually rather than by autopilot control increases the odds of a failure by a factor of two. The more flights the organization has in total, and the more flights on a given type of airframe, the lower the failure rate. More flights on a given tail number increases the risk of a failure.

The results of this research were obtained from data gathered on small, unmanned aerial systems with wingspans between 20 inches and 11 feet and takeoff weights under 100 pounds. Twenty nine unique airframes (with a total of 103 different tail numbers), including a mix of lab-designed prototypes and COTS models were aggregated for this analysis. All data were obtained in a research environment where prototype SUAS are frequently developed and more traditional, COTS SUAS are flown in new ways and with novel objectives, payloads, and technologies. Thus, the results of this research are applicable primarily to experimental vehicles and in a research and development environment. This is not to say that the lessons learned must not be applied to other systems or operational environments, but merely that one should exercise caution and be fully aware of the underlying assumptions of this research before applying its conclusions to other scenarios.

Recommendations

The recommendations from these results are fairly straightforward. One simple way to decrease the odds of a failure is to substitute a COTS SUAS for a prototype SUAS whenever possible. This should be done especially when flying high value payloads or

mission-critical objectives. Alternately, (and much more complexly) the prototype aircraft could be brought up to COTS levels of reliability. However, given that AFRL is primarily tasked with pushing the technological boundaries and then transferring the technology to other organizations or private industry to be refined, this second option is outside the normal scope of operations and is almost certainly not cost effective.

The preference for autonomous flight over manual flight to reduce failure rates is not necessarily intuitive but makes sense in light of the remarkable differences in sensory environment that SUAS exhibit versus manned aircraft. The possibilities for perceptual errors have been well-established for large UAS. It appears that autonomous control of SUAS significantly reduces failures that would have otherwise occurred with manual flight.

Less significant, but still important is the role of experience in failure prevention. Greater organizational experience, expressed as the increase in total number of flights across all platforms, reduces failures. Greater organizational experience with a given type of airframe similarly reduces failure rates. These results were largely expected, but are nonspecific given the quality of the data. The term "experience" is not merely a measure of AFRL's proficiency with the mechanics of SUAS flight tests and knowledge of the peculiarities specific to each airframe. Rather, this broad term incorporates all organizational knowledge and improvements made to SUAS operations and airframes without identifying the specific improvements that reduced the failure rates. AFRL continually adds additional features to its flight planning and operations and iterates on SUAS designs to great overall effect. The result has been a statistically significant decrease in failure rates over time, which can be captured in this concept of

organizational experience. However, the specific improvements that have had greatest effect (or those hypothesized improvements which have actually worsened failure rates) cannot be determined with precision from the data. Thus, while increased organizational experience with flight testing and on each airframe is likely to continue to lower failure rates, the efficacy of specific actions and policy decisions have not been assessed in the research, except to the extent that they influence other variables.

One such case of this influence is with pilot currency. While AFRL was not required to meet mandatory pilot currency requirements for the period covered by this data, future regulations will incorporate requirements mandating that SUAS pilots have a minimum number of flights over a set time period, in order to remain “current”. The results of this research indicate that pilot currency is not statistically significant in the model of SUAS failures. Coupled with the results on the benefits of autonomous flight over manual flight, it appears that resources would be best spent to ensure that the autopilot settings are correct rather than that pilots have recently flown. The elimination of a pilot currency requirement, while not impacting failure rates, would also save valuable range testing time that can be used for higher priority flight experiments.

Many recent AFRL flight experiment test plans have imposed maximum surface wind requirements (which were not in place while this data was being collected) that can cause test delays or cancellations. This research demonstrated that the maximum surface winds at the test site were statistically insignificant in the model of SUAS failures. Likewise, other environmental factors such as time of day, temperature, and location had no statistical impact on failure rates. Thus, there is not enough evidence to conclude that

any of these measured environmental factors impact SUAS failure rates either positively or negatively.

Flight failures have historically occurred in 38.5% of all AFRL/RWWV SUAS flight experiments. An understanding of this failure rate may help decision makers, range safety officers and test engineers with expectation management. While this research has outlined some positive steps AFRL can take to lower mishap rates, it has also identified areas that show little promise at improving the rates in the hopes that preventive measures are only undertaken which are statistically justifiable and whose benefits are appropriately balanced with costs. There are a few additional measures that can be taken that may assist future analysts and engineers identify means to further lower SUAS failure rates.

Root causes of failures should be analyzed from an engineering perspective and tracked to identify trends. This could be as simple as one or two lines added to every flight report and one or more categories assigned to the outcome of each flight in a database, much like the error codes of the DoD HFACS taxonomy. This simple addition will enable a future analyst to quickly identify failure or damage trends without resorting to guesswork or memory to recall the root causes. Additionally, if any other factors that were not included in this research are deemed important for possible failure prediction and prevention (such as percent of maximum takeoff weight used, ground station operator experience, or mission type as a categorical variable), they should be recorded in the flight reports.

Tracking each tail number individually would help to identify trends in aircraft disposal for reliability estimates. Each tail number was not tracked precisely over the five

years covered by this data set. We do not know with certainty from the data where each tail number went: whether it was scrapped when a program ended, disposed of following a crash, upgraded into a newer airframe model, demolished in destructive lab testing, or sent away to be a desk model or display aircraft. By tracking the outcomes of flight testing on each airframe, reliability estimates may be made that can shed light on how many flights an airframe can be expected to have before being disposed, or what the mean time between failures is for a tail number.

Lastly, organizational experience has a positive impact on failure rates, but there were insufficient data to determine with specificity which changes were beneficial and which were detrimental. Over the period measured, the net result was improvement in failure rates, but there is no way to identify and quantify the most cost-effective improvements. A record of policy decisions, major design alterations, or major process changes should be noted on flight reports to provide a time stamp for future analysis. This future analysis should seek to determine whether the policy, design, or process changes have been effective in lowering failure rates, predicting damage rates, or generally improving the cost-effectiveness of operations.

Areas for Future Research

Ordinarily, a designed experiment is recommended to better screen important features and to optimize SUAS failure rates. Unfortunately, no designed experiment is possible in this case. This is due to the unique nature of the data; only a handful of parameters can be adjusted to specific factor levels, while most cannot. The surface wind

speed can be measured but not controlled. Test times, days and locations are typically awarded on a priority-based system, with no guarantees of dates, times or locations. Number of total flights can never be lowered, and can only be raised incrementally. The same is true of the other counting variables. Each airframe has a given number of flights in its history and can only gain them at the cost of adding an additional airframe flight, an additional organizational flight, and an additional pilot flight, while at the same time resetting all the days since last flight, days since last mission, days since pilot's last flight and similar interval measures. While techniques like analysis of covariance (ANCOVA) could be used to account for the influence of uncontrollable factors, the interconnected nature of the flights precludes a randomized, designed experiment on the full complement of parameters.

Any effects of selection bias should be investigated. The results of this research describe significant correlations that were found in the data, but these correlations do not necessarily imply causation. For example, the logistic regression failure model found that as the number of flights on a tail number increases, its odds of a failure increase. This does not necessarily mean that tail numbers should be scrapped after a few flights to lower their risk of failure. It could mean that older aircraft are intentionally selected for riskier flight experiments – nothing in the data is able to identify if that hypothesized action is occurring. Likewise, the fact that wind speed did not affect failure rates should not be read as an encouragement to fly in adverse weather conditions. It could be that only missions with higher-likelihoods of success (as determined by the test engineer) were selected for known windy days, or that other tests were intentionally scrapped despite the lack of maximum wind regulations at the time. These examples highlight how

selection bias could influence the results and should be investigated to better characterize the effectiveness of potential interventions.

Contributions of this Research

- The first published study of SUAS failure and damage rates, this research quantified the risk of data loss associated with SUAS flight test failures and the probability of damage incurred during flight testing.

- Analyzed 20 measurable parameters and identified both statistically significant and insignificant factors that affected SUAS failure rates.

- Developed and validated a logistic regression model to predict the probability of a flight failure and to quantify the increased or decreased risk associated with alternate flight test configurations.

- Developed a model to predict the minimum number of SUAS flights necessary to achieve any specified level of expected mission success.

- Proposed targeted and statistically justifiable failure prevention techniques to be implemented by test engineers and decision makers to reduce the risk of data loss associated with SUAS flight testing.

Bibliography

Abate, Gregg, Kelly Stewart, and Judson Babcock. *Autonomous Aerodynamic Control of Micro Air Vehicles.* Eglin AFB, FL: Air Force Research Laboratory, 2009.

AFRL. "AFRL FLIGHT TEST AND EVALUATION." *AFRLMAN 99-103.* USAF, 2007.

—. "AFRL RESEARCH TEST MANAGEMENT." *AFRLI 61-103.* USAF, 2007.

Agresti, Alan. *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons, 2002.

Allison, Paul. "Missing Data." Chap. 4 in *The SAGE Handbook of Quantitative Methods in Psychology*, edited by Roger Milsap and Alberto Maydeu-Olivares, 72-89. Thousand Oaks, CA: SAGE Publications, Inc, 2009.

Ballesteros, Jose Sanchez-Alarcos. *Improving Air Safety through Organizational Learning.* Hampshire: Ashgate, 2007.

Bauer, Kenneth. "Course Notes: OPER685." *Applied Multivariate Analysis I.* Presented at AFIT, Spring Quarter, 2011.

Bauer, Kenneth, Stephen Alsing, and Kelly Greene. "Feature screening using signal-to-noise ratios." *Neurocomputing* 31 (2000): 29-44.

Carney, Ryan, Rodney Walker, and Peter Corke. "Image Processing Algorithms for UAV "Sense and Avoid"." *Proceedings of the IEEE International Conference on robotics and Automation.* Orlando, FL: IEEE, 2006. 2848-2853.

Chen, Xiang-Hui, Arthur Dempster, and Jun Liu. "Weighted Finite Population Sampling to Maximize Entropy." *Biometrika* 81, no. 3 (1994): 457-69.

Cline, Charles. *Methods for Improvements in Airworthiness of Small UAS.* North Carolina State University, 2008.

Clothier, Reece, Jennifer Palmer, Rodney Walker, and Neale Fulton. "Definition of an Airworthiness Certification Framework for Civil Unmanned Aircraft Systems." *Safety Science* 49 (2011): 871-885.

Dalamagkidis, K, K.P. Valavanis, and L.A. Piegl. "On Unmanned Aicraft Systems Issues, Challenges and Operational Restrictions Preventing Integration into the National Airspace System." *Progress in Aerospace Sciences* 44 (2008): 503-519.

Dekker, Sidney. "Illusions of Explanation: A Critical Essay on Error Classification." *The International Journal of Aviation Psychology* 3, no. 2 (2003): 95 - 106.

Dermentzoudis, Marinos. *Establishment of Models and Data Tracking for Small UAV Reliability.* Naval Postgraduate School, 2004.

DoD. *Department of Defense Human Factors Guide.* DoD, 2005.

Gibb, Randy. *Classification of Air Force Aviation Accidents: Mishap Trends and Prevention.* Report CI04-1814, US Air Force, 2006.

Haimes, Yacov. *Risk Modeling, Assessment, and Management.* Third Edition. Hoboken, NJ: John Wiley & Sons, 2009.

Hawkins, Scott, and Reid Hastie. "Hindsight: Biased Judgments of Past Events After the Outcomes Are Known." *Psychological Bulletin* 107 (1990): 311-327.

Hosmer, David, and Stanley Lemeshow. *Applied Logistic Regression.* New York: John Wiley & Sons, 1989.

Johnson, Chris. "Act in Haste, Repent at Leisure: An Overview of Operational Incidents Involving UAVs in Afghanistan (2003-2005)." *3rd IET Systems Safety Conference.* Birmingham, UK, 2008.

Joint Aviation Authorities. *The human factors implications for flight safety of recent developments in the airline industry.* Flight Safety Foundation, 2003.

Jones, Bruce, Michelle Canham-Chervak, and David Sleet. "An Evidence-Based Public Health Approach to Injury Priorities and Prevention." *American Journal of Preventive Medicine* 38, no. 1S (2010): S1-S10.

Magister, Tone. "The Small Unmanned Aircraft Blunt Criterion Based Injury Potential Estimation." *Safety Science* 48 (2010): 1313-1320.

Mejias, Luis, Daniel Fitzgerald, Pillar Eng, and Xi Liu. "Forced Landing Technologies for Unmanned Aerial Vehicles: Towards Safer Operations." *Aerial Vehicles*, 2009: 415-442.

Menard, Scott. *Applied Logistic Regression Analysis.* Second. Thousand Oaks, CA: Sage Publications, 2002.

Miarecki, Sandra, and Stefan Constable. *An Assessment of Thermal Stress Effects on Flight Mishaps that Involve Pilot Human Factors.* Brooks City-Base, TX: Performance Enhancement Directorate, 2007.

Nullmeyer, Robert, Robert Herz, and Gregg Montijo. "Training Interventions to Reduce Air Force Predator Mishaps." *15th International Symposium on Aviation Psychology.* Dayton, OH, 2009.

OSD. *UAV Roadmap 2009-2034.* OSD, 2009.

OSD. *Unmanned Aerial Vehicle Reliability Study.* OSD, 2003.

Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies.* Princeton, NJ: Princeton University Press, 1999.

Reason, James. *Human Error.* New York: Cambridge University Press, 1990.

Salas, Eduardo, C. Burke, Clint Bowers, and Katherine Wilson. "Team Training in the Skies: Does Crew Resource Management (CRM) Work?" *Human Factors* 43, no. 4 (2001): 641-674.

Schreiber, Brian, Don Lyon, Elizabeth Martin, and Herk Confer. *Impact of Prior Flight Experience on Learning Predator UAV Operator Skills.* Mesa, AZ: Air Force Research Lab, 2002.

Sharma, Sanjiv, and D Chakravarti. "UAV Operations: An Analysis of Incidents and Accidents with Human Factors and Crew Resource Management Perspective." *Indian Journal of Aerospace Medicine* 49, no. 1 (2005).

Steppe, Jean. *Feature and Model Selection in Feedforward Neural Networks.* Air Force Institute of Technology, 1994.

Tarr, Gregory. *Multi-Layered Feedforward Neural Networks for Image Segmentation.* Air Force Institute of Technology, 1991.

Teets, Edward, Casey Donahue, Ken Underwood, and Jeffrey Bauer. *Atmospheric Considerations for Uninhabited Aerial Vehicle (UAV) Flight Test Planning.* Edwards, CA: NASA, 1998.

Tvaryanas, Anthony, and William Thompson. "Recurrent Error Pathways in HFACS Data: Analysis of 95 Mishaps with Remotely Piloted Aircraft." *Aviation, Space, and Environmental Medicine* 79, no. 5 (May 2008): 525 - 532.

Tvaryanas, Anthony, and William Thompson. *Unmanned Aircraft System (UAS) Operator Error Mishaps: An Evidence-based Prioritization of Human Factors Issues.* DoD, 2006.

Tvaryanas, Anthony, William Thompson, and Stefan Constable. "Human factors in Remotely Piloted Aircraft Operations: HFACS Analysis of 221 Mishaps Over 10 Years." *Aviation, Space and Environmental Medicine* 77, no. 7 (2006): 724 - 732.

Van Houten, John S. *Forecasting Aircraft Mishaps Using Monthly Maintenance Reports.* Monterrey, CA: Naval Post Graduate School, 1994.

Wang, Y. H. "On the Number of Success in Independent Trials." *Statistica Sinica* 3 (1993): 295-312.

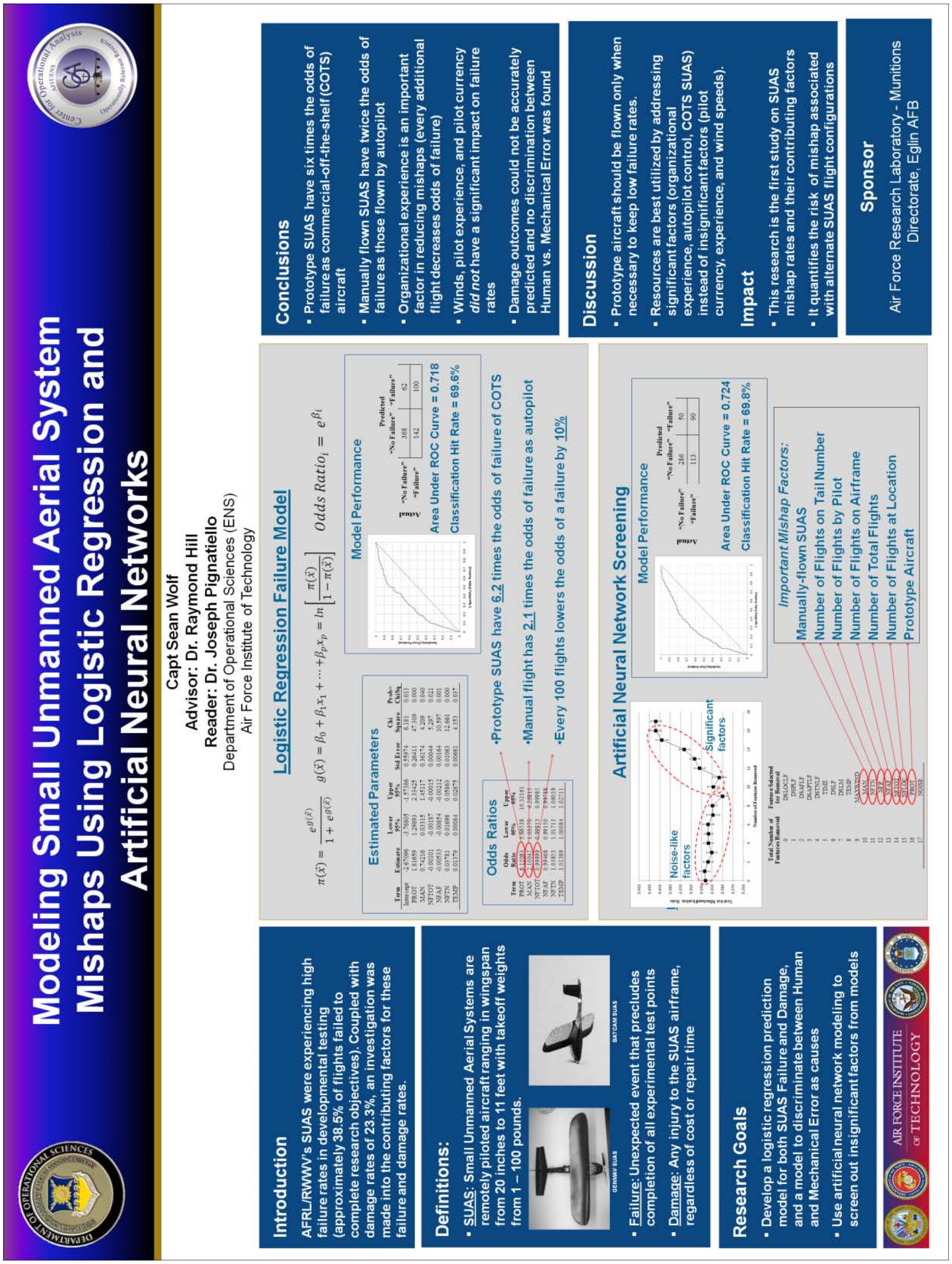
Weibel, Ronald, and John Hansman. *Safety Considerations for Operation of Unmanned Aerial Vehicles in the National Airspace System.* Cambridge, MA: MIT International Center for Air Transportation, 2005.

Wiegmann, Douglas, and Scott Shappell. *A Human Error Approach to Aviation Accident Analysis.* Burlington, VT: Ashgate, 2003.

Williams, Kevin. *A Summary of Unmanned Aircraft/Incident Data: Human Factors Implications.* US Department of Transportation, Federal Aviation Administration, 2004.

Williams, Kevin. *Unmanned Aircraft Pilot Medical and Certification Requirements.* Oklahoma City, OK: Federal Aviation Administration, 2007.

Appendix: Storyboard Slide



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
17-02-2012	**Master's Thesis**	Aug 2010 – Mar 2012

4. TITLE AND SUBTITLE

MODELING SMALL UNMANNED AERIAL SYSTEM MISHAPS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wolf, Sean E., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 641
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-OR-MS-ENS-12-29

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RWWV
Attn: Mr. Johnny Evers
101 W Eglin Blvd — COM: (850) 882-8876
Eglin AFB FL 32542 — e-mail: Johnny.Evers@eglin.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A dataset of 854 small unmanned aerial system (SUAS) flight experiments from 2005-2009 is analyzed to determine significant factors that contribute to mishaps. The data from 29 airframes of different designs and technology readiness levels were aggregated. 20 measured parameters from each flight experiment are investigated, including wind speed, pilot experience, number of prior flights, pilot currency, etc. Outcomes of failures (loss of flight data) and damage (injury to airframe) are classified by logistic regression modeling and artificial neural network analysis.

From the analysis, it can be concluded that SUAS damage is a random event that cannot be predicted with greater accuracy than guessing. Failures can be predicted with greater accuracy (38.5% occurrence, model hit rate 69.6%). Five significant factors were identified by both the neural networks and logistic regression.

SUAS prototypes risk failures at six times the odds of their commercially manufactured counterparts. Likewise, manually controlled SUAS have twice the odds of experiencing a failure as those autonomously controlled. Wind speeds, pilot experience, and pilot currency were not found to be statistically significant to flight outcomes. The implications of these results for decision makers, range safety officers and test engineers are discussed.

15. SUBJECT TERMS

Small Unmanned Aerial Systems, SUAS, Micro Air Vehicles, MAV, Unmanned Aerial Vehicles, UAV, Aviation Mishaps, Flight Safety, Mishap Rates, Logistic Regression Modeling, Artificial Neural Networks, Feature Screening

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	U	117	Dr. Raymond Hill, AFIT/ENS (937) 255-3636, ext 7469; e-mail: Raymond.Hill@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18